Neurophysiological and computational bases of goal-directed behavior

Ruggero Basanisi

NNT/NL : 0000AIXM0000/000ED000

Ph.D. THESIS

Defended at Aix-Marseille Université
on 17th December 2021 by

Ruggero Basanisi

Neurophysiological and computational bases of goal-directed behavior

Discipline
Biologie santé

Spécialité
Neurosciences

École doctorale
ED62

Laboratoire/Partenaires de recherche
Institut de Neurosciences de la Timone

Composition du jury

- Andrea Brovelli — Directeur de thèse
  Institut de Neurosciences de la Timone
- Sylvia Wirth — Rapporteur
  Institut des Sciences Cognitives
- Mehdi Khamassi — Rapporteur
  Institut des Systèmes Intelligents et de Robotique
- Boris Burle — Examinateur
  Laboratoire de Neurosciences Cognitives
- Christian Bénar — Examinateur
  Institut de Neurosciences des Systèmes
- Marc Deffains — Examinateur
  Institut des Maladies Neurodégénératives

Acknowledgments

I would like to start by thanking my two supervisors Andrea Brovelli and Paul Apicella for their mentoring and their friendship. We shared several good moments together. You are good mentors from a scientific and human point of view, and I feel really lucky to have worked with you. A great thanks to GIanluca Baldassarre, for his teaching, his perseverance, and the time he dedicated to me before and during my PhD. I see you as a source of inspiration. Then I would like to thank my partner in life, Michela, who always supports me, even after a long time! Thanks to share together all the moments, from the happiest to the saddest, during relaxed or stressful periods. On that note, yes we need a vacation! Thanks to my family, who always supported me even when the communication on skype was hard. I would like to thank all the BraiNets team, and especially, I will thank Etienne BG, for his friendship, his patience, the long discussions, the scientific debates, the dead zombies, the good advice, the jokes, the St. Nectaire cheese, all the cutted onions… and all the moments spent together. A big thanks to Michele, for the dinners, the nights out and for answering to all (and they were a lot) my questions about physics and mathematics. A big thanks also to two of the most important people I met here in Marseille, Pavlosky and Lorensky. If you ever read these lines, I know exactly what you thought/said when you stopped at the previous sentence point: oooh yeah. That's creepy, but also kinda cool. I want to thanks all the colleagues of the INT: Jean Charles, for having welcomed us warmly in his family; la coloc, Melina, Alex and Charlie (and also Lulu/Mimou), for all the good times, the laughter and the good dinner together; Davidou (YES), I don't remember why… as I told before that's creepy, but also kinda cool; Hannah, epic psych-running (haloo); Sandra la profesora (pero no pudo repetir nada de las clases de español aquí), Lucio il bello, Antoine, Maud, Manon… and really you are a lot, if you ever shared a beer or a laugh with me, thanks. Another big thanks to all the friends with whom I spent most of my

weekends hanging around, playing music, going to calanques, etc..: Romanos, Julien, Ivan, Carla, Fatima, Antony, Chiara, Marwa, Jonas, Cem, Luc, Leda, Nicolas, Martina, Nicola, Fabiana, Thais... I should stop this, surely I forgot someone please forgive me. Thanks to my old distant friends, in particular to Livia, Daniel, Marta, Naomi and Sami, I know I should call you more often. As a wise man said “uno piacere, ha il grande piacere, ospite di avere ospite”. This will be a personal commitment.

Index of content

Acknowledgments 1

Index of content 3

Abstract (en) 6

Abstract (fr) 8

List of studies* 10

Section 1. Introduction 11

1.1 Behavioral principles of instrumental learning 11

1.1.1 Classical or Pavlovian learning 12

1.1.2 Instrumental learning 14

Experimental paradigms for instrumental learning 14

Pavlovian to instrumental transfer (PIT) 16

Goal-directed learning 16

Habits 18

Experimental paradigms to determine if a behavior is goal-directed 19

1.1.3 Toward a unified vision of learning 20

1.2 Brain circuits of goal-directed learning 21

1.2.1 Fronto-striatal loops 22

1.2.2 Cortical regions 23

Orbitofrontal cortex (OFC) 25

Ventromedial prefrontal cortex (vmPFC) 25

Dorsolateral prefrontal cortex (dlPFC) 26

Posterior parietal cortex (PPC) 27

1.2.3 Subcortical regions 27

Striatum 28

Amygdala 30

Hippocampus 30

1.3 Advantages and pitfalls of brain data acquisition techniques for the study of goal-directed learning 31

1.3.1 Spikes and Local field potentials (LFPs) 32
1.3.2 Electro- and Magneto-encephalography (EEG and MEG) 34
1.3.3 Complementary techniques 39
1.4 Computational models of goal-directed learning 40
1.4.1 Neural networks 41
A glimpse of history 41
Spiking Neural Networks 43
1.4.2 Reinforcement learning models (RLM) 43
HB-GDB modulation: the arbitrator model 45
1.4.3 Bayesian models for goal-directed causal learning 47
1.5 Identifying the neural correlates of goal-directed learning 48
1.5.1 Model-free and model-based analysis of brain data 49
1.5.2 Information theory 50
Mutual Information (MI) 52
1.6 Thesis objectives 53
1.7 Publications 54

Section 2. A generative spiking neural-network model of goal-directed behaviour and one-step planning 55

Section 3. Beta oscillations in the monkey striatum encodes reward prediction error 88

Section 4. Dynamics of human cortical circuits mediating goal-directed causal learning 122

Section 5. Neuroinformatics, tools, Open Science 155
5.1 Team resources 155
5.2 Softwares development 155
5.3 Open science 156
5.4 NeuroMatch Academy - Deep Learning 157

Section 6. Discussion and Conclusion 158
6.1 A generative spiking neural-network model of goal-directed behaviour and one-step planning - Faced problems and solutions to GDB modelling 159
6.2 Beta oscillations in the monkey striatum encodes reward prediction error - The role of beta-band oscillations in striatal RPE signaling 163

6.3 Dynamics of human cortical circuits mediating goal-directed causal learning - High-gamma activity in human prefrontal cortex reflects relevant behavioral aspects of goal-directed causal learning 166

6.4 Future perspectives 170

Bibliography **172**

Abstract (en)

In the field of instrumental learning, mammals are able to implement two different behavioral strategies to interact with the environment: goal directed behavior (GDB), computationally flexible but slow, suitable to learn new tasks and adapt to changing environments; and habitual behavior, hard-coded, but suitable for faster motor responses and facing recurrent tasks. The advantage of GDB resides in the use of an inner representation of the environment, a 'model of the world', to encode stimuli-actions-outcomes associations, and its exploitation to choose future actions, in a process called planning. GDB is supported by large-scale networks involving both cortical and subcortical regions. Nevertheless, several open questions still remain. The aim of this thesis is to contribute to the understanding of three open questions (declined in three studies) that pertain to the neural and computational mechanisms of GDB.

In the first study, we investigated how complex computations, such as learning the model of the world and planning, can emerge from simple neural activity. To achieve that, we built a spiking neural network, able to encode stimulus-actions-outcomes associations as a hidden Markov model (HMM), using biologically inspired mechanisms such as spike-timing dependent plasticity (STDP), and to test this model to correctly plan actions in order to solve a visuomotor goal directed task. The performance of the model was validated on behavioral data from human participants that performed the same task.

In the second study, we assessed the importance of striatum in encoding the reward prediction error (RPE) signals, a relevant update signal in most instrumental learning models. To do so, we analysed local field potentials (LFPs) recorded in rhesus macaque striatum while performing a probabilistic goal-directed learning task. Then, we computed the trial-by-trial RPE using a Q-learning model fitted on monkeys' behavior. Our results showed a significant increase of mutual information

(MI) between the beta-band (15-30Hz) oscillatory activity and the RPE after the outcome presentation. Moreover, such correlates of RPE signals form an anatomo-functional gradient in the striatum, showing stronger effects toward the rostro-ventral part and vanishing toward the caudo-dorsal part.

In the third study, we investigated the neural correlates of GDB at the whole-brain cortical level in humans. To do so, we recorded the brain activity of human participants using magnetoencephalography (MEG) while they were performing a goal-directed causal learning task. We exploited cortical high-gamma activity (HGA, 60-120Hz) to map the spatio-temporal dynamics during learning. In particular, we used an ideal observer Bayesian model to estimate the trial-by-trial evolution of relevant behavioral variables, such as action-outcome probabilities and contingency values. We used MI and group-level cluster-based statics between HGA and those variables to obtain a whole brain profile of behavioral-dependent regions of interests' activity, confirming some results from the literature.

Abstract (fr)

Dans le domaine de l'apprentissage instrumental, les mammifères sont capables de mettre en œuvre deux stratégies comportementales différentes pour interagir avec l'environnement: le comportement dirigé vers un but (“goal-directed behavior”, GDB), flexible sur le plan computationnel mais lent, adapté à l'apprentissage de nouvelles tâches et à l'adaptation à des environnements changeants; et le comportement habituel, encodé de façon rigide, mais adapté à des réponses motrices plus rapide, adapté aux tâches récurrentes. L'avantage du GDB réside dans l'utilisation d'une représentation interne de l'environnement, un ‘modèle du monde’, pour encoder les associations stimuli-actions-conséquences, et dans l'utilisation de ce modèle pour choisir les actions futures au cours du processus de planification. Le GDB est soutenu par des réseaux cérébraux à grande échelle impliquant des régions corticales et sous-corticales. Néanmoins, plusieurs questions ouvertes demeurent. L'objectif de cette thèse est de contribuer à la compréhension de trois questions ouvertes (déclinées en trois études) qui concernent les mécanismes neuronaux et computationnels du GDB.

Dans une première étude, nous avons cherché à savoir comment des calculs complexes, tels que l'apprentissage du modèle du monde et la planification, peuvent émerger de l’activité neuronale. Pour ce faire, nous avons construit un réseau de neurones actifs, capable d'encoder des associations stimulus-actions-conséquences sous la forme d'un modèle de Markov caché (Hidden Markov Model, HMM), en utilisant des mécanismes d'inspiration biologique tels que la ‘spike-timing dependent plasticity’ (STDP), et d'utiliser ce modèle pour planifier correctement des actions afin de résoudre une tâche visuomotrice. Les performances du modèle ont été validées sur des données comportementales de participants humains ayant effectué la même tâche.

Dans une deuxième étude, nous avons évalué l'importance du striatum dans l'encodage de l'erreur de prédiction de la récompense (Reward Prediction Error, RPE), un signal de mise à jour pertinent dans la plupart des modèles d'apprentissage instrumental. Pour ce faire, nous avons analysé les potentiels de champ locaux (Local Field Potentials, LFP) enregistrés dans le striatum de macaques rhésus pendant l'exécution d'une tâche d'apprentissage probabiliste dirigée vers un but. Ensuite, nous avons calculé la RPE essai par essai en utilisant un modèle de 'Q-learning' adapté au comportement des singes. Nos résultats ont montré une augmentation significative de l'information mutuelle (Mutual Information, MI) entre l'activité oscillatoire dans la bande bêta (15-30 Hz) et la RPE après le résultat de l'action. De plus l'information sur la RPE forme un gradient impliquant l'ensemble du striatum, plus intense dans la partie rostro-ventrale que dans la partie caudo-dorsale.

Dans la troisième étude, nous avons étudié les corrélats neuronaux du GDB au niveau cortical du cerveau entier chez l'homme. Pour ce faire, nous avons enregistré l'activité corticale de participants humains à l'aide de la magnétoencéphalographie (MEG) pendant qu'ils effectuaient une tâche d'apprentissage causal dirigée vers un but. Nous nous sommes concentrés sur l'extraction et l'analyse de l'activité oscillatoire dans la bande gamma haute (High-Gamma Activity, HGA 60-120 Hz) pour mapper la dynamique spatio-temporelle pendant l'apprentissage. Ensuite, nous avons utilisé un modèle Bayésien d'observateur idéal pour estimer l'évolution essai par essai des variables comportementales pertinentes, telles que les probabilités de résultats d'action et les valeurs de contingence. Nous avons utilisé la MI et des statiques au niveau du groupe basées sur le cluster entre le HGA et ces variables pour obtenir un profil du cerveau entier de l'activité des régions d'intérêt dépendant du comportement, confirmant certains résultats de la littérature.

List of studies*

Section 2**: A generative spiking neural-network model of goal-directed behaviour and one-step planning.
By: Basanisi Ruggero, Brovelli Andrea, Cartoni Emilio, Baldassarre Gianluca (2020)
(Peer reviewed article published on *PLoS Computational Biology*)

Section 3: Beta oscillations in the monkey striatum encodes reward prediction error.
By: Basanisi Ruggero, Marche Kevin, Combrisson Etienne, Apicella Paul, Brovelli Andrea (2021)
(Article in preparation)

Section 4: Dynamics of human cortical circuits mediating goal-directed causal learning.
By: Basanisi Ruggero, Combrisson Etienne, Dauce Emmanuel, Brovelli Andrea (2021)
(Article in preparation)

N.B.: *A complete list of publications can be found in **Section 1.7**. Each of these studies has its own bibliography, listed at the end of each corresponding section. The bibliography at the end of this manuscript only refers to the remaining sections.
The study described in **Section 2 started during an Erasmus stage in Marseille, to then be refined and finalised during my PhD. The contribution of Gianluca Baldassarre in this project was fundamental for its good success.

Section 1. Introduction

1.1 Behavioral principles of instrumental learning

In order to cope with a constantly changing environment, animals are faced with the complex task of rapidly adapting to a huge variety of incoming stimuli, perceived through different sensory channels, and to select the most appropriate behaviors. One of the most basic forms of motor response are reflexes. Reflex actions are automatic, involuntary and fast; they are triggered by a sensory stimulus and supported by the so-called reflex arcs, the neural pathways controlling a reflex. For example, if we touch a hot surface, the withdrawal reflex allows us to retract our hand before we can have serious injuries (Hultborn, 2006). Interestingly, although reflexes do not involve directly the central nervous system, they can be modulated by descending signals from the brain. For example, during prepulse inhibition (PPI) experimental paradigm, a stimulus (a pre-pulse) inhibits a startle response consequent to an aversive acoustic or tactile stimulus (Li et al., 2009). On the other hand, in order to perform more complex behavioural responses and plan multiple actions, an agent should be able to combine information about external and internal stimuli, as well as past experiences and predictions about future outcomes. A cognitive function that supports the ability to acquire and integrate knowledge about the relations among stimuli, actions and outcomes is associative learning (Shanks, 1995; Wasserman and Miller, 1997; Mitchell et al., 2009; Dickinson, 2012). Associative learning is the ability to learn contingency relations between events in their environment (De Houwer, 2009) and it reflects a fundamental component of adaptive behavior. Two large categories of associative learning are classically defined: classical conditioning, describing stimulus-stimulus associations learning; and instrumental learning, describing stimulus-action and action-outcome

associations learning. I will now briefly outline the basic principles of classical conditioning, and then focus on instrumental learning, which is the main topic of my PhD. project.

1.1.1 Classical or Pavlovian learning

In behavioural psychology, two distinct classes of learning are acknowledged: classical learning and instrumental learning. Classical or Pavlovian conditioning, formulated by Ivan Petrovic Pavlov in 1927 (Pavlov and Anrep, 1927), is defined as the ability to learn stimuli-stimuli associations. According to this paradigm, if the experimenter presents a salient stimulus, as for example some food, to a dog, he will show a salivary response anticipating the consummatory act (if the dog has previous knowledge about the smell and the appearance of that food, and if he already consumed it before). Thus, food is defined as the unconditioned stimulus (US), while the subsequent salivation takes the name of unconditioned response (UR). The pairing of a second non-salient stimulus, as for example the ring of a bell, together with the US for a sufficient number of times, will lead to the creation of an "association", which will suffice to trigger the salivary response, even in the absence of the food. In this case, the ring of the bell takes the name of conditioned stimulus (CS), because it is associated with the US, while the salivary response takes the name of conditioned response (CR). Importantly, one of the first observations made on this paradigm was that a prominent factor for the conditioning to happen, is the timing of the occurrence of the non-salient stimulus during learning, called contiguity. Indeed, the conditioning will be effective only if the CS and the US are contiguous in time and space, meaning that the CS should be presented before, during, or shortly after the presentation of the US, but at the same time it should not occur too early or too late. Thus, we can define a precise effective temporal window for learning to occur.

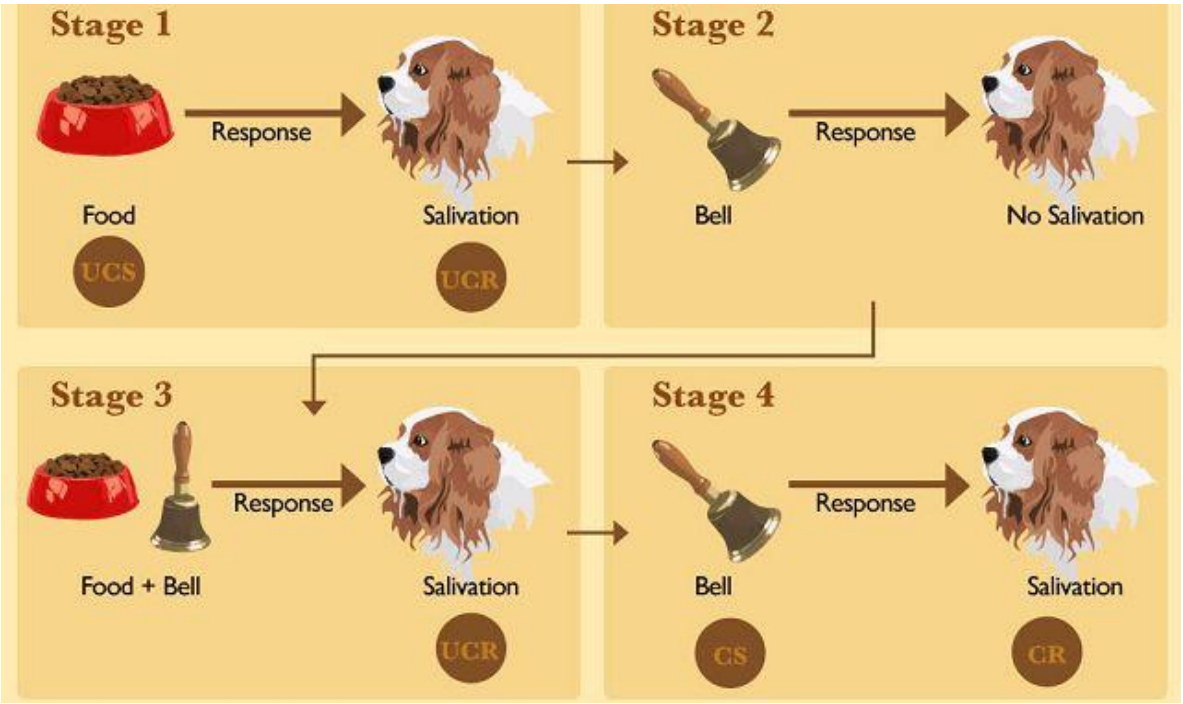


Figure 1. *The Pavlovian paradigm for classical learning. Image taken on Google.*

The paradigm of classical conditioning can be generalised to aversive stimuli (i.e., negative reinforcers), that is all the stimuli that normally induce an aversive UR, and that are thus associated with a negative value; they are stimuli that an agent would try to avoid. Another form of classical learning is multiple-order conditioning (Rizley and Rescorla, 1972), which is defined as the ability to associate a second non-salient stimulus, such as the lighting-up of a light bulb, to the previously associated CS (ringing bell). Such pairing will trigger salivation as a second-order CR. Finally, a last aspect of classical conditioning extensively studied in associative learning is extinction (Skinner, 1938). Extinction refers to the gradual decrease in response to a conditioned stimulus that occurs when the stimulus is presented without reinforcement. For example, once the experimenter creates one or multiple-order CS, if they're presented repeatedly without giving the opportunity to then carry out the consummatory behavior, the association between CS and CR will slowly vanish. Interestingly, if after extinction the experimenter wants the CS to take back its salience, the dog will need less training time to restore the association between CS and CR. It is worth noting that classical conditioning does not rely on

the motivational state, indeed it is possible to use the same paradigm with an aversive stimulus, that will be followed by an aversive response.

1.1.2 Instrumental learning

Classical conditioning does not depend on the actions performed by the agent. On the other hand, in instrumental or operant conditioning, the US depends on instrumental behavior. In other words, the agent is asked to respond with a voluntary behaviour that can be triggered or inhibited by a reinforcement (reward or punishment). In this framework, a stimulus can be used to signal the subject about the possibility to perform the motor response in order to achieve the desired result; but an explicit stimulus (e.g., a light or a tone) is not always needed, indeed it can be represented by the current ensemble of environmental stimuli, also called context. Thus, instrumental learning is a type of associative learning process through which the strength of a behavior is modified by reinforcement or punishment.

Experimental paradigms for instrumental learning

Edward Lee Thorndike was one of the first scientists to describe instrumental behavior with a simple experimental paradigm leading to the development of operant conditioning within Behaviorism. His major contribution to the field consisted in a novel approach to quantify the behavioral changes occuring during instrumental learning. He used a cage with an opening mechanism that could be activated through pulling a rope, and he placed a cat inside of it and a visible reward outside of it. The cat learned by trial-and-error to pull the rope, in order to open the cage and earn the reward, and showed with training a reduction of the time required to perform the task (Thorndike, 1898). The decrease in reaction times was considered as an index for learning. Thanks to this paradigm, Thorndike formulated the ‘law of effect’, stating that motor responses that produce a pleasing effect in a particular context become more likely to occur again in that context, while motor

responses that produce an unpleasant effect become less likely to occur again in that context.

Another classical experimental setup to study instrumental learning, invented by Burrhus Frederic Skinner in 1938 (Skinner, 1938), is the so-called Skinner box (**Figure 2**). It is composed of different signaling devices, such as a speaker and some lights, some input response devices such as levers, and a reward/punishment delivery device such as a pellet dispenser or an electrified grid. The Skinner box allowed for the first time the training of animals with as little as possible intervention and the development of quantifiable training protocols, the so-called reinforcement schedules, that manipulate learning by means of varying ratios (VR) or rewards or variable intervals (VI) of time between rewards. With this set-up, it is possible, for example, to teach an agent to press a lever in response to a visual stimulus in order to receive a pellet unit or to avoid an electric shock. Different variations of this simple example allow to dissociate different aspects of the behavior. In this case, the definition about the positive or negative value subjectively attributed to the reinforcer does not depend on what can be considered pleasant or unpleasant, but on the empirical rate of performed motor responses to obtain or avoid the reinforcer. Importantly, this paradigm led to development of the modern reinforcement learning (RL) theory, where a reward or a punishment can act respectively as positive or negative reinforcer of the learned instrumental behavior (Sutton and Barto, 1998).

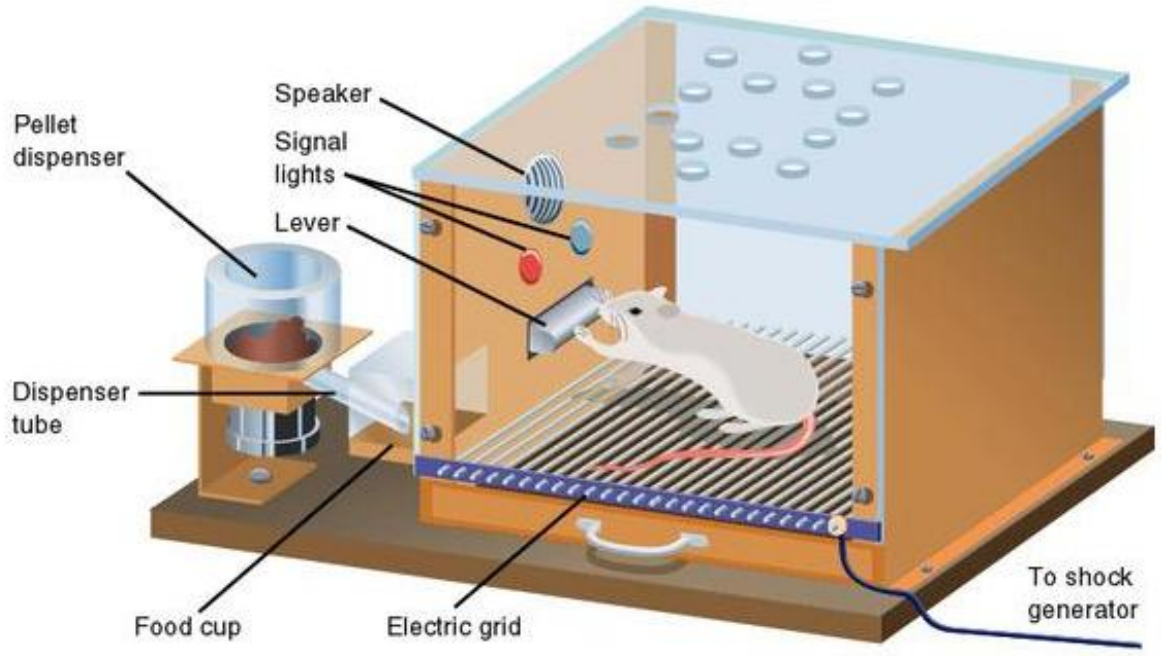


Figure 2. *The Skinner box and its principal components. Image taken on Google.*

Pavlovian to instrumental transfer (PIT)

Although classical and instrumental conditioning belong to separate categories of associative learning, they nevertheless share some common properties that can be highlighted in experimental paradigms, such as the 'Pavlovian to instrumental transfer' (PIT), according to which after conditioning a subject on a stimulus-reward association (e.g. sound-pellet) and on an action-reward association (e.g. lever pressing-reward), the stimulus will be able to trigger the action, meaning that a stimulus-action (sound-lever) association is formed (for a very exhaustive explanation see (Holmes et al., 2010)).

Goal-directed learning

Modern associative learning theories suggest that instrumental behaviors are controlled by complementary, but interacting, systems that lead to different behavioral strategies: goal-directed and habitual learning (Dickinson and Balleine, 1994, 2000; Keramati et al., 2011; Dolan and Dayan, 2013).

Goal-directed learning is driven by internal goals and motivational state, it is flexible and used in particular to find solutions to new problems or to face changing conditions. In general, goal-directed behaviors are computationally intense and not suitable to provide a fast and automatic motor response. Goals are defined differently from rewards or action's outcome. Goals are the starting point of the willful control of actions (Gollwitzer and Moskowitz), they are ideal desired states that drive behavior, in which one or a set of conditions are satisfied (as for example in a reward maximization task) (Ressler, 2004). During the acquisition of a GDB, the associations between actions and their outcomes are learned.

The acquisition of goal-directed behaviors leads to creation of internal representation of contingencies between actions and outcomes (Blaisdell, 2006; Penn and Povinelli, 2007; Liljeholm, 2018, 2021). Indeed, goal-directed learning can be defined as the ability to learn if a certain action can effectively cause or prevent a given outcome, or if actually there is no association between them. Goal-directed learning therefore forms the basis of a key cognitive function, which supports the creation of our sense of causality between our behaviors and their outcomes. According to a popular model in cognitive psychology, the sense of causality can be quantified experimentally as the action-outcome contingency, called ΔP. The action-outcome contingency is defined as the difference between two conditional probabilities: $P(O|A)$, that is the probability associated to the outcome when the agent perform an action; and $P(O|\neg A)$, that is the probability that the outcome spontaneously occurs not associated with the agent's action (Allan, 1980; Allan and Jenkins, 1980; Hammond, 1980; Allan, 1993; Allan et al., 2008; Morris et al., 2017). Being the result of the difference between two probabilities, ΔP can take all the values from -1 to 1; positive values are associated with a positive sense of causality (the action triggers the outcome), negative values are associated with a negative sense of causality (the action prevents the outcome), while values close to zero gives no sense of causality (the action and the outcome have no causal dependence).

These actions-outcomes contingency values can be learned by trial-and-error, retained in memory, and used to adjust behavior with respect to changing context, providing an efficient strategy to reach the goal. Thus, GDB is defined as model-based (Lee et al., 2014), meaning that it uses an internal representation of the world and transition probabilities between actions and outcomes. This resonates with the notion of a cognitive map (Tolman, 1948) that keeps track of previous experiences in order to orient future actions. Indeed, an additional component of goal-directed learning is the ability to plan future behaviors according to internal goals and motivational states. In other words, planning is the ability to use the knowledge about the model of the world, in order to program future actions. Since knowledge about the structure of the environment is collected from the interaction with the environment, such information about the experienced actions-outcomes is kept into memory, and then used to efficiently explore the environment. Thus, another key feature of planning is to use a still partially observed model of the world to select actions (Bonet and Geffner, 2014).

Habits

The second form of behavioral strategy supporting instrumental learning concerns habits. Habitual behaviors are inflexible, and arise from long-term training and consolidation of stimulus-response-outcome associations. Habits can arise in particular to respond to familiar problems or to face well-known tasks, thus it results computationally light and suitable to provide a fast motor response whether possible (Balleine and O'Doherty, 2010). Habitual behavior is outcome independent, indeed to trigger an HB it is sufficient that the subject perceives a stimulus that is strongly enough associated with an action to produce the motor response. Habits are thought to be primarily triggered by antecedent stimuli, rather than the prediction of future outcomes. For this reason, habits are normally considered as model-free: there is no need for the agent to keep track of the actions or the actions-outcomes transition probabilities or the internal representation of the task

(Graybiel, 2008; Dolan and Dayan, 2013). The insensibility to the outcomes and the absence of a model, and therefore of planning, are the causes of HB's inflexibility and speed. Indeed, if we train an agent for a long enough time on a task, consequently to a slow consolidation he will learn a HB, and from then on it will be very hard for him to change his behavior also if the rules of the task changes (Yin and Knowlton, 2006; Hilario, 2008). To modify his behavior, the agent will need a very long time. On the other hand, the agent will be very good in performing the original task for which he developed a habit, always giving the correct answer in a short time, as soon as a stimulus appears.

Experimental paradigms to determine if a behavior is goal-directed

There exists two main experimental paradigms that can be used to establish if an observed behavior can be considered as goal-directed or not: 1) outcome devaluation and 2) contingency degradation. Outcome devaluation paradigm was defined by Dickinson in 1985 (Dickinson, 1985) and refined by Balleine and Dickinson in 1998 (Balleine and Dickinson, 1998). The aim of outcome devaluation is to assess if the behavior of an agent changes accordingly with changes in the value assigned to an outcome. According to this paradigm, we can train an agent to perform two different actions each one leading to a different type of outcome (e.g. different food), at the beginning the agent will perform the two actions equally across time. Then, it is possible to devalue one of the two outcomes, for example making it always available; after that, if the behavior is goal-directed, the agent should lose interest in performing the motor response leading to the devalued outcome, to fully focus on the other action. In this case, if the agent had established a habit, he would have continued performing both the actions with the same frequency.

Contingency degradation was at first observed by Robert A. Rescorla (Rescorla, 1966, 1968) on studies about classical learning: after an agent was trained to respond to a CS, if the corresponding US was presented without being preceded by the CS for

enough times, the response to the CS decreases over time. This paradigm was then extended to instrumental behavior (Adams and Dickinson, 1981; Schreiner et al., 2020). As an example, we can train an agent to perform an action in order to receive a desired outcome. After training, if we start to give him that outcome at some random point in time, also if he doesn't perform the action, the agent can lose interest in performing the trained action. This would be linked to the fact that its sense of action-outcome contingency, or causal sensation, will drop close to zero, meaning that his behavior was still goal-directed as he updated its internal model. On the contrary, if the agent established a HB, he would have continued to perform the action independently of the introduced devaluation, at least for a long period of time, until he comes back to a goal-directed strategy.

Although the transition from GDB to HB and vice-versa are well described phenomena, less is still known on how an agent is called to perform an action in a goal-directed or in a habitual way, and which is the computational mechanism underlying this switch. So far, one of the most accredited hypotheses is the existence of a cognitive computational arbitrator model that selects one of the two behaviors. I will give a better overview of these arbitrator models in **Section 1.4.3**.

1.1.3 Toward a unified vision of learning

As we can see, there is some similitude in between Pavlovian conditioning and habitual behavior. A recent review proposes a slight modification of this cognitive organization, proposing a dichotomous division between a stimulus-driven model-free control and a goal-directed model-based control (Corbetta and Shulman, 2002; O'Doherty et al., 2017). The first category comprehends the reflexes, the Pavlovian classical conditioning and the HB. Those three types of learning share indeed some characteristics such as the fact that they take control over the actions in a rapid and efficient way, the fact that they are automatically deployed, inflexible and hard to modify, moreover they are model-free and outcome independent.

Stimulus-driven control can be thought of as *retrospective*, in that it depends on integrating past experiences. The second category comprehends exclusively the GDB, whose main characteristics are: slow but flexible computations, the need of a model, and the fact that it can be stimulus independent. Goal-directed control may be thought of as *prospective* in that it leverages a cognitive map of the decision problem to flexibly reevaluate states and action. Overall, open questions exist concerning the relation between habit and goal-directed learning, and a unified theory is still missing.

1.2 Brain circuits of goal-directed learning

Instrumental learning is thought to be mediated by the activity of neural circuits and populations distributed over fronto-striatal loops (O'Doherty et al. 2017).
Instrumental learning has been extensively investigated with the use of both animal models and in humans, highlighting the role of cortical and subcortical areas involved in its implementation. Each of this network can be composed of several cortical and subcortical functionally connected brain regions, able to express different cognitive aspects of the behavior (**Figure 3**).
In goal-directed behavior, one of the most important concepts is the idea that an agent needs to be able to represent the values of the outcomes, in order to build an efficient cognitive map that allows him to compute how to achieve the desired outcome, that is the one with the highest value, and thus the sequences of actions that can lead him to that. In order to do so, the brain should also be able to establish associations between states, whether they represent stimuli, actions or outcome. In this section I will give a general overview of the participation of different brain regions, at first cortical and then subcortical, and of their interactions in instrumental learning.

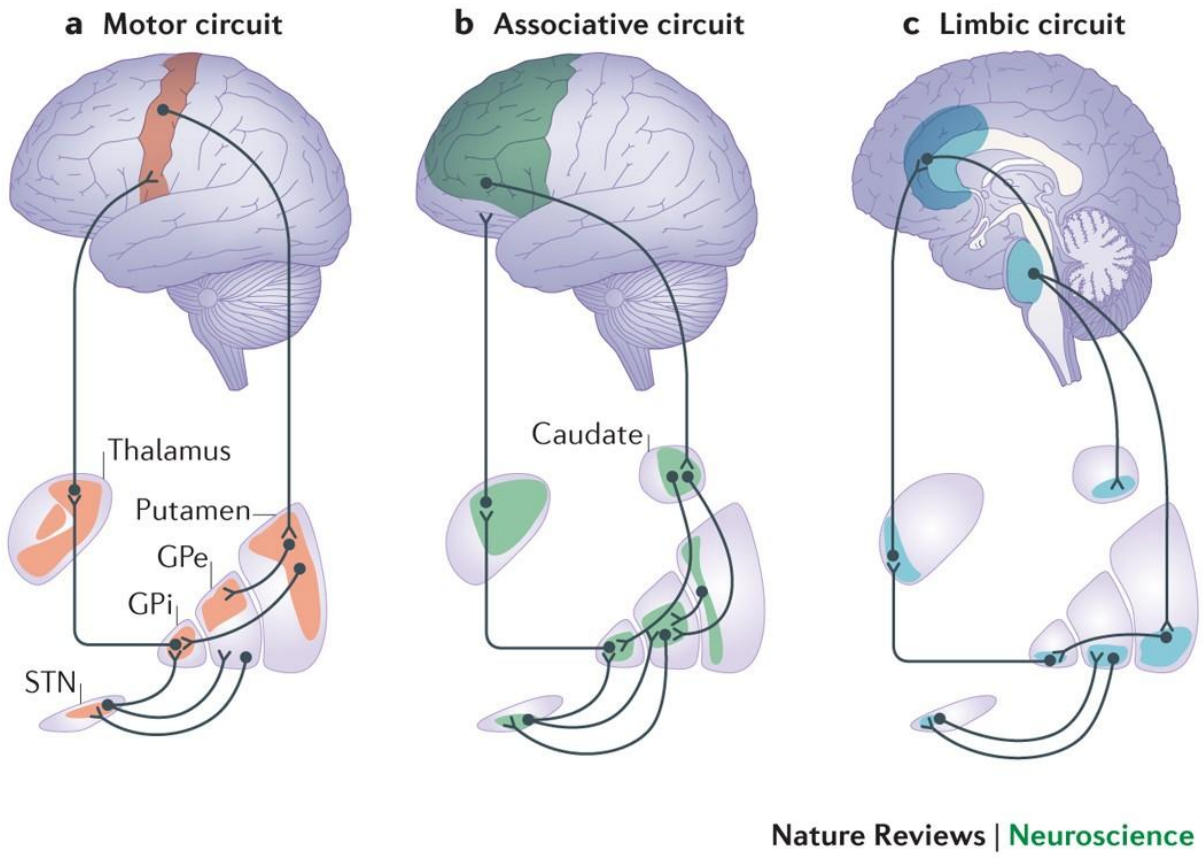


Figure 3. *Schematization of the three fronto-striatal loops. Image taken from Jahanshahi et al., (2015).*

1.2.1 Fronto-striatal loops

Physiological, anatomical and imaging studies in both human and non human primates, revealed that the basal ganglia complex follows an intrinsic anatomo-functional organization, forming cortico-basal ganglia loops of connections implied in different aspects of behavioral control involving different cortical regions (Haber, 2003; Nakano et al., 2000; Redgrave et al., 2010; Liljeholm and O'Doherty, 2012; Jahanshahi et al., 2015; Morris et al., 2016). Three main distinct fronto-striatal loops are identified: 1) the limbic loop, implied in motivational and emotional aspects, that involves the ventral part of the striatum, the anterior cingulate cortex, the orbitofrontal cortex and the amygdala; 2) the associative loop, implied in planning and higher cognitive control, involving the

anterior part of the striatum (dorso-medial striatum in rodents), the dlPFC and the PPC; 3) the sensorimotor loop, implied in motor control, that involves the posterior part of the striatum striatum (dorso-lateral striatum in rodents) and the sensorimotor and supplementary motor cortices. Also if the differences among these loops are well identified, the anatomy of these circuits doesn't follow a strict separation, but more a transitional gradient (**Figure 4**) (Vogelsang and D'Esposito, 2018; Han et al., 2021). Some studies revealed that the transition from goal-directed to habitual behavior can rely on a gradual switching between the fronto-striatal loops, especially from the associative to the sensorimotor networks (Yin and Knowlton, 2006; Ashby et al., 2010). Overall, these studies suggest that goal-directed learning is based on the associative and limbic fronto-striatal circuits.

1.2.2 Cortical regions

Before going into details, a small clarification here is needed: usually when we talk about cortical regions we talk about regions belonging to the neocortex, and thus to the frontal, parietal, temporal and occipital lobe. The neocortex has a very conservative structure, made of six layers, each one containing the bodies of different cellular types, organised in cortical columns (core sections perpendicular to the brain surface of about half a millimeter diameter, comprising all the six layers). The thickness of each layer can vary depending on the location and the function of that part of the cortex we are considering; a very well known example of that is the primary motor cortex, in which layer IV (inner granular layer) is thinner in favour of a thicker layer V (inner pyramidal projection neurons layer). The layers, and thus, the cortical columns are all oriented on the same axis, with the axons of the neurons perpendicular to the surface, and are organised in convolutions that form sulci, in order to maximise the surface on volume ratio. But the neocortex represents 90% of the whole cortex, the remaining 10% is represented by the allocortex, which only has 3-4 layers, and which comprehends olfactory and limbic

structures, such as the insular pole and the hippocampus. This clarification is due because in **Study 3**, we will use marsatlas for the brain parcelization of human participants; this atlas includes the insular pole among the cortical regions (together with neocortical structures), and the hippocampus among subcortical regions (together with the basal ganglia, the thalamus and other limbic structures such as the amygdala, also if it belong to cortex), but that is just for labeling simplicity. For the sake of uniformity, I will follow the same subdivision also in the following paragraph in which I will outline the current hypotheses regarding the role of different cortical areas in goal-directed learning.

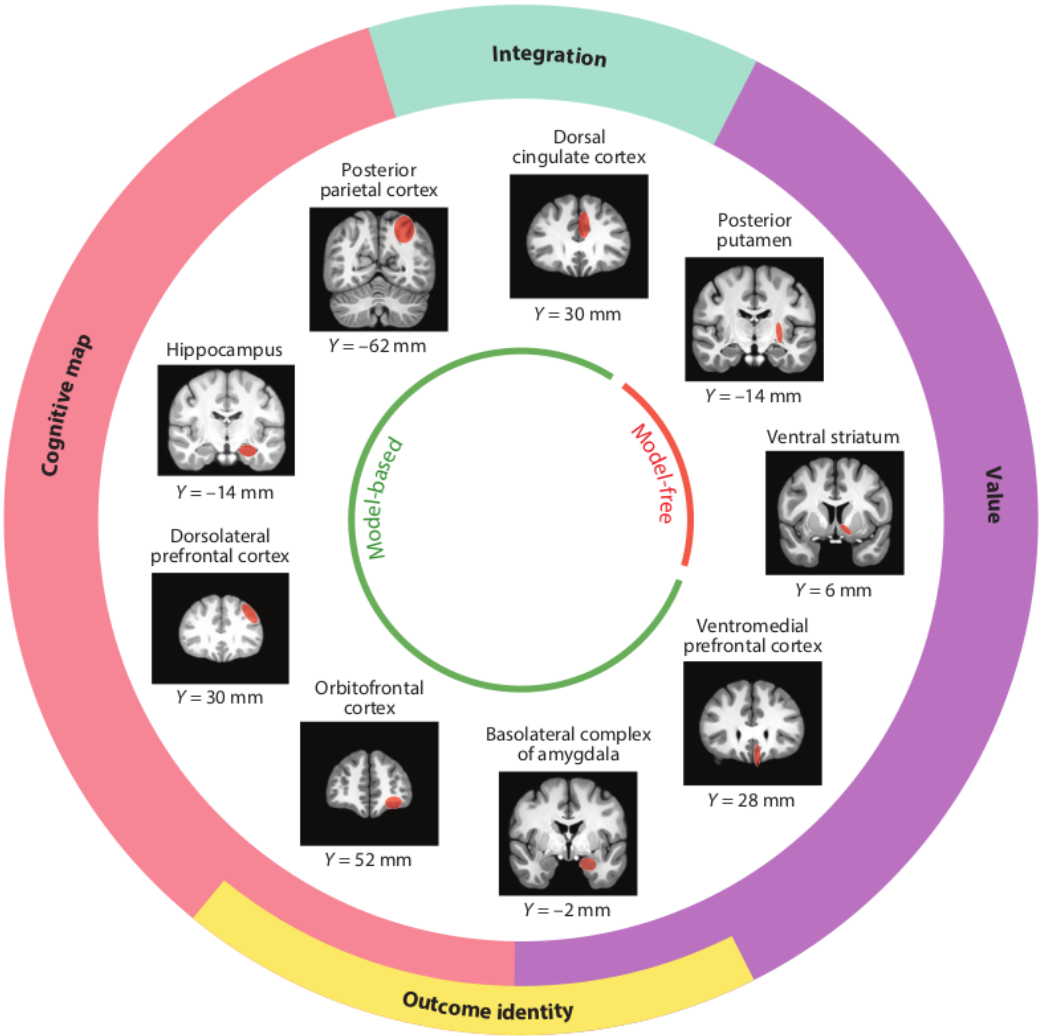


Figure 4. *Cortical and subcortical areas involved in different aspects of instrumental learning. Image taken from O'Doherty et al., (2017).*

Orbitofrontal cortex (OFC)

OFC has several key roles in instrumental learning, spanning from encoding the cognitive map to the representation of outcomes' identity and expected and effective value. Regarding its role in encoding the cognitive map, computational studies suggest that the OFC is able to represent states, in particular in an abstract task space (Wilson et al., 2014). The OFC encodes preferentially stimuli and outcomes associations instead of actions. Indeed, it is able to encode for expected value based on a stimulus-stimulus association, and to encode the outcome identity, activating in presence of stimuli which predicts those outcomes (Howard et al., 2015). Moreover, the OFC seems to differently respond to the values of conditioned stimuli to unconditioned appetitive or aversive stimuli, and to the predicted values of those conditioned stimuli (Schoenbaum et al., 1998; Salzman et al., 2007; Salzman and Fusi, 2010). Other studies showed that OFC discriminates between different amounts of values of the outcomes, and the values of expected and prospective outcomes (Padoa-Schioppa and Assad, 2006; McDannald et al., 2011). OFC responds also to different kinds of already experienced outcomes and responds differently according to the motivational state associated with them (O'Doherty et al., 2001; Rolls, 2003; Smith et al., 2010). The role of OFC in goal-directed learning is still object of intense studying (for a deepening see: https://psycnet.apa.org/PsycARTICLES/journal/bne/135/2)

Ventromedial prefrontal cortex (vmPFC)

vmPFC is involved mostly in representing outcomes' value, and shares some functions with OFC, such as responding accordingly with the amount of value attributed to an outcome, encoding outcomes value after their reception, and econding motivational value assigned to outcomes (O'Doherty et al., 2001; Rolls, 2003; Padoa-Schioppa and Assad, 2006; Smith et al., 2010). Pan et al. in 2014 observed that monkeys' lateral prefrontal cortex can compute higher-order outcomes values, indeed, its activation correlates with the value of an outcome

associated to a novel stimuli, and inferred by the previously experienced stimuli-outcomes associations; the human vmPFC seems to act in a similar way (O'Doherty et al., 2017). Another study demonstrated that the activity of vmPFC scales with the outcome values, responding with an increase in activity for positive values and a decrease in activity for negative values (Plassmann et al., 2010). Sometimes an agent is called to evaluate different types of outcome and to compare them, vmPFC seems to be involved in assigning a common currency to different outcome's categories to allow a comparison (Chib et al., 2009; Levy and Glimcher, 2012). Also, vmPFC seems able to encode the incentive value of the actions, and the action-outcome causal relation in an instrumental contingency learning task (Matsumoto et al., 2003; Liljeholm et al., 2011).Moreover a recent studies indicates its role in positive reward associated prediction errors (Gueguen et al., 2021).

Dorsolateral prefrontal cortex (dlPFC)

dlPFC is related to the ability of building cognitive maps involving actions, and in action planning (Balleine and Dickinson, 1998). In a 2010 paper, Glascher and colleagues proposed the existence of a state prediction error (SPE), another type of prediction error, not based on reward, that acts like a signal to update model-based expectations, which measure the surprise of a new state based on the current estimate of the state-action-state transition probability (Gläscher et al., 2010). Using fMRI, they found out that dlPFC correlates with SPE, meaning that this region can be involved in learning cognitive models that involve actions. Moreover, in order to build an internal model that takes in consideration the actions, the agent should be able also to retain in his memory the past actions and the transitions between states, and dlPFC is indeed associated with working memory (Levy and Goldman-Rakic, 2000; Miller and Cohen, 2001; Procyk and Goldman-Rakic, 2006).

Posterior parietal cortex (PPC)

PPC covers several different aspects of decision making. It participates, for example, in perceptual decision making, that is the ability to establish the identity of a stimulus in a limited space of categories, useful for state identification (Shadlen and Newsome, 2001). According to those findings, other studies showed that PPC encodes the category of current or future potential states and stimuli (Freedman and Assad, 2006; Doll et al., 2015). The activity in the inferior parietal lobule, a part of the PPC, has been found to vary according to the causal contingency measure resulting as a function of two outcome probabilities, called ΔP, together with actions rates and judgment of the causal efficacy of those actions (Liljeholm et al., 2011, 2013). Moreover, as dlPFC, PPC was found to respond to SPE (Gläscher et al., 2010) and to participate in action values representation and action planning, as it has a well established role in numerical cognition (Platt and Glimcher, 1999).

1.2.3 Subcortical regions

Subcortical brain regions comprehend a variety of different structures, all with different roles, essential for sustaining higher cortical computations. A major complex is represented by the basal ganglia, an ensemble of nuclei that was first thought to contribute mostly to motor functions, and was later found to be involved in higher cognitive processes and emotions (Lanciego et al., 2012). The basal ganglia complex includes:

- **striatum**: it is the main component, a very complex structure at the connectivity, cellular and molecular level, that in primates is subdivided in a ventral part containing the nucleus accumbens (NAc) and a dorsal part, including two nuclei, the putamen and the caudate nucleus. From this structure originates both the direct and the indirect basal ganglia pathways;
- **globus pallidus**: structure composed by GABAergic neurons that receives GABAergic afferents from the striatum, it can be subdivided in its external

portion (GPe) that projects on subthalamic nucleus participating in the indirect pathway, and its internal portion (GPi) that projects on the thalamus participating in the direct pathway;

- **subthalamic nucleus** (STN): small nucleus that is intensively studied for its clinical relevance, especially after the advent of deep brain stimulation (DBS), an effective treatment to reduce symptoms in Parkinson's disease. It is also involved in the hyperdirect basal ganglia pathway, receiving excitatory inputs directly from the cortex and sending its projections toward GPi and the substantia nigra pars reticulata.
- **substantia nigra**, that is subdivided in two parts: the pars reticulata (SNr) that receives GABAergic afferents from striatum and GPe nucleus and glutamatergic afferents from the STN, and sends GABAergic efferent projections to the thalamus; the pars compacta (SNc) that receives GABAergic afferents from the striatum and sends modulatory dopaminergic efferent projections to the striatum together with the ventral tegmental area (VTA).

Striatum

The whole striatum receives glutamatergic projections from the cortex and the thalamus (called corticostriatal and thalamostriatal projections respectively) and receives midbrain dopaminergic projections from the SNc and VTA. The SNc and VTA are two regions that are well known to be involved in the encoding of reward value and reward prediction error (RPE) (Apicella et al., 1991; Schultz, 2016a, 2016b). Moreover, it receives afferences from the amygdala and the hippocampus. It sends GABAergic projections to the GPi and GPe (called striatopallidal projections) forming respectively the direct and the indirect striatal pathways. The activation tuning of those two pathways is allowed by the same nature of the dopamine and of its receptors. Indeed, the effect of dopamine on D1 receptors is to activate GABAergic neurons involved in the direct pathway, inhibiting the GABAergic neurons of the GPi and SNr, that results in a thalamic activation; while dopamine

inhibits neurons expressing D2 receptors, allowing GPe GABAergic neurons involved in the indirect pathway to be activated, thus inhibiting the STN, that sends glutamatergic excitatory efferent projections to the GPi and SNr, allowing them to inhibit the thalamus. The synergy between these two pathways allows the fine regulation of the thalamocortical circuits controlling behavioral expression. The three main nuclei of the striatum (NAc, putamen, and caudate) are associated with different functions. The ventral striatum, to which we refer as the 'limbic striatum' for its implication in the limbic loop, has been implicated in reward circuit and encoding of RPE signals, as shown notably by human fMRI studies (Delgado et al., 2005; Wang et al., 2016). Moreover it is implied in motivational aspects and decision making. Ventral striatum sends projections to the GPi and SNr, and contacts cortices associated with limbic functions such as the rostral cingulate cortices. NAc is also involved in predictions of CS linked to both appetitive and aversive US, and in conditioned skeletomotor reflexes such as consummatory or avoidance responses (O'Doherty et al., 2017). Dorsal striatum (caudate and putamen) is instead implied in motor functions and in their inhibitory control, in stimulus-action associations learning (Balleine et al., 2007; Bissonette and Roesch, 2015; Yager et al., 2015) and in punishment (Pessiglione et al., 2006; Palminteri et al., 2012; Palminteri and Pessiglione, 2017).

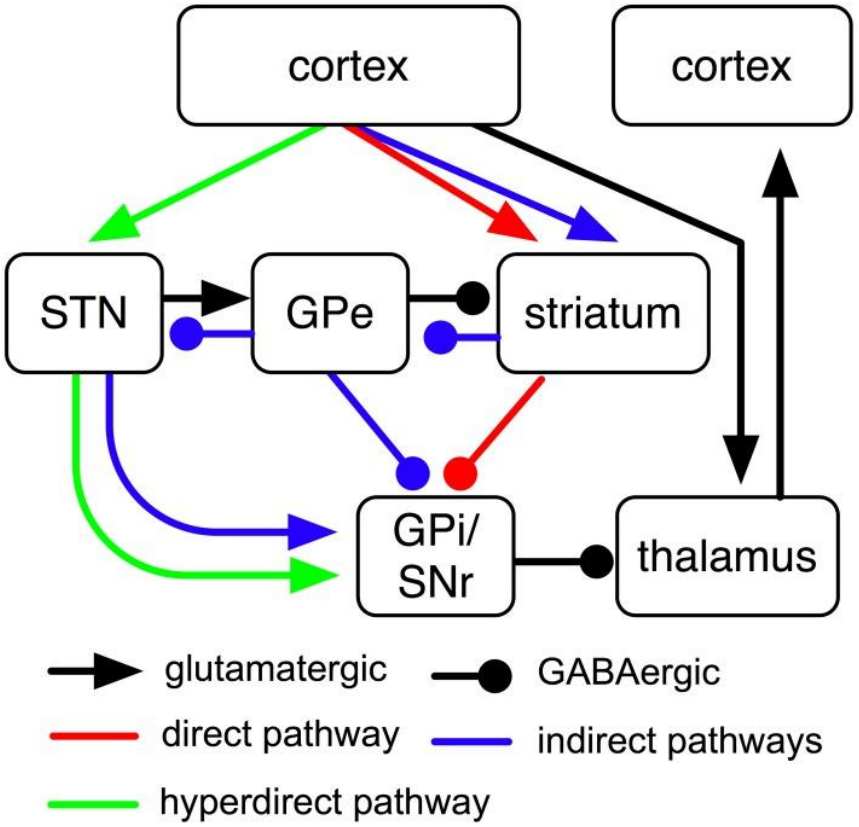


Figure 5. *Schematization of the striatal pathways. Image taken from Schroll and Hamker, (2013).*

Amygdala

Amygdala (Amy) is a nucleus belonging to the limbic complex, and thus involved in emotions, such as fear, and in some rapid behavioral response like the fight or flight response, or conditioned automatic reflexes (LeDoux et al., 1988). Studies in rodents and monkeys showed that Amy is also involved in encoding conditioned stimuli when they are linked to unconditioned appetitive or aversive stimuli, moreover it is involved in representing context, stimulus identity, and reward expectation (Schoenbaum et al., 1998; Paton et al., 2006; Salzman and Fusi, 2010).

Hippocampus

Hippocampus (Hipp) belongs to the allocortex and can be subdivided in 5 parts: cornus ammonis (CA) from 1 (more external, in continuity with the subiculum) to 4

(more internal) and the dentate gyrus (DG). Hipp has a very well established role in declarative long-term memory and in representing space through place cells, able to inform the agent about a specific position in space, but not following a specific pattern as grid cells in the entorhinal cortex, with whom they communicate (Bird and Burgess, 2008; Moser et al., 2008). For those reasons, Hipp was always considered a good candidate to encode cognitive maps, especially in spatial decision making tasks where a model-based planning is needed, and indeed the activity of place cells can represent the agent trajectory during a spatial decision-making task (Pfeiffer and Foster, 2013). Hipp seems to be more involved in stimuli-stimuli associations encoding, more than actions-outcomes associations, as some study showed its ability to link reward to perceived stimuli (Wimmer and Shohamy, 2012).

To conclude, current literature suggests that goal-directed learning is supported by subcortical areas, through the expression of rostro-caudal gradients involving the basal ganglia and the cortical brain regions.

1.3 Advantages and pitfalls of brain data acquisition techniques for the study of goal-directed learning

Learning is surely a brain network phenomenon. On the other hand, functional specificity exists at the microscopic and mesoscopic level. One of the challenges of future studies will be to integrate brain data from multiple spatial and temporal scales so as to have a complete picture of the neural bases of goal-directed learning. This section introduces the state-of-the-art concerning the methodological approaches for the analysis of the neural correlates of goal-directed learning and the underlying computationsValid correlates for neural activity find their roots in different kinds of signals, electrical, biological, optical and so forth; the only limit is

engineers' fantasy, and technological or computational limitations. Each of these techniques has its own pros and cons that should be taken in consideration during experimental design. The constraints that are taken in consideration are spatial resolution, temporal resolution, mobility and coverage.

Evidently, there exists no single experimental technique that allows the measurement of brain activity at both a high spatial and temporal resolution in humans or non-human primate. In this section I will describe two of the techniques that I exploited for the studies described in **Section 3** (LFP) and **Section 4** (MEG). Moreover, I will compare them with similar techniques, and finally I will give a brief description of other data acquisition techniques.

1.3.1 Spikes and Local field potentials (LFPs)

Neurons are cells able to transmit information to each other using electrical and chemical signals. When a neuron generates an electrical impulse, that is called action potential or spike, it is transmitted through the axon to then reach the synapses, which, releasing neurotransmitters, generate a postsynaptic potential (PSP) in the dendrites of adjacent neurons.

Using microelectrodes, we are able to record the spiking activity of single neurons in the brain of behaving subjects, using the difference of potential with respect to a reference. The electrical signal is usually band pass filtered between 300 and 6000 Hz in order to capture just the fastest events. It is possible to perform spikes detection during (by hardware) or after (by software) the recording in order to obtain a time series that can be used to compute, for example, the inter-spikes interval or the firing rate of that neuron. A relevant feature of spikes is that it is possible to understand which kind of neuron we are recording by observing the spike waveform.

On the contrary, Local Field Potentials (LFPs) represent larger and slower electrical phenomena, recorded in a radius of 0.5 - 2 millimeters from the tip of the electrode and low-pass filtered with a cutoff frequency in the range of 100-300 Hz (Buzsáki,

2006). About their origin, LFPs are thought to be the result of the synchronization of the synaptic potentials (both excitatory and inhibitory postsynaptic potentials (E/IPSP), and sometimes also membrane hyperpolarization) occurring in that radius (Buzsáki, 2006; van der Meer, 2010; Buzsáki et al., 2012). LFPs are particularly useful to study oscillatory activity, indeed after preprocessing and artifact rejection, we can use the time series to extract the power of several frequency bands in order to build a time-frequency map that describes how synchronous activity evolves for each frequency band in time (preprocessing and power extraction techniques will be better described in the next paragraph). Spikes and LFPs recordings can be combined to compute the spike-LFP phase-coupling, which is especially effective for the study of long range interactions.

The aim of all of these approaches is to relate the properly treated signal with some relevant behavioral variable recorded at the same time as the neurophysiological signal, in order to further proceed with descriptive analysis and statistics.

Those two techniques have the advantage of a very high temporal resolution, but on the other side they are very invasive, meaning that they need a surgical procedure in order for the electrode to be placed. This leads to disadvantages, like the fact that the experimenter should be careful in inserting the electrodes, especially if he is supposed to reach a deep part of the brain, and the fact that the recording position is relatively unknown. Indeed a common procedure is to make use of anatomical atlas and stereotaxic coordinates to implant the electrodes, to then cause an electrical or thermal damage before removing the electrodes, in order to verify their recording site in a postmortem histological analysis. However, the brain coverage of those techniques is getting better and better, from single pin electrodes we passed to multiple-pins electrodes and microelectrodes (e.g. NeuroPixels), arrays of electrodes ad microelectrodes (e.g. Utah Arrays), and recently even to record an entire hemisphere of a behaving macaque monkey, using a large-scale semi-chronic microdrive recording system developed in Charles Gray's lab (Dotson et al., 2017).

In **Section 3** I analysed a dataset recorded by the team of Paul Apicella to investigate the neural correlates of goal-directed learning in behaving non-human primates striatum, performing a free-choice goal-directed learning task. In particular, I wanted to assess how striatal oscillatory activity correlates with relevant learning signals such as RPE in different striatal fields.

1.3.2 Electro- and Magneto-encephalography (EEG and MEG)

Most popular for human studies, EEG and MEG are non-invasive whole-brain recording techniques. They differ in the acquisition phase, but they share very similar data analysis pipelines.

In an EEG, a soft plastic or silicon cap containing several equidistant holes is placed on the head of a subject. This cap is used to hold in place the EEG electrodes: the experimenter injects inside each hole, on the subject's scalp, some conductive gel to then place the electrodes on the top of it. The origin of the signal is the same as the LFP one and it's recorded as an electric field in the order of microvolts (mV, 10^{-3}V).

In MEG, a hard plastic cap, often associated with a chair or a table, is placed on the subject's head. This cap already contains the sensors in a fixed position, the sensor can be of two types: magnetometers, to measure the magnetic field, or gradiometers, which are pairs of magnetometers placed very close one each other to measure the difference in magnetic field between them. MEG machine needs a couple more attentions compared to EEG setting: the machine should be isolated from magnetic fields with the use of a metal alloy called mu-metal, that has infinite magnetic permeability; moreover the coils used in MEG sensors should be able to record magnetic field in the order of femtotesla (fT, 10^{-15}T), thus they need to be constantly kept under very low temperature using liquid helium. The origin of the magnetic field is attributed to synchronous excitatory or inhibitory PSPs of several

close neurons, acting like small electrical wires that generate a magnetic field orthogonal to current direction. Thus, the magnetic field is perpendicular to neuronal axon direction, and in fact its power is maximal in the correspondence of cortical sulci's walls, and minimal on sulci's ridges.

Both in EEG and in MEG it is possible to add electrodes out of the scalp surface to record eye movements (vertical, horizontal and blinks) or cardiac activity: those signals will be used during preprocessing to remove artifacts.

The result of an EEG or MEG recording is composed of an ensemble of neurophysiological time series at the sensor level, that at first should pass through preprocessing. This stage is used to clean the data from artifacts and includes different steps such as: notch filtering (a band-stop filter used to subtract the periodic influence of electrical current from the signal, that is 50Hz in Europe), artifact rejection by independent or principal component analysis (ICA or PCA), band-pass filtering (or high-pass or low-pass), and a crucial visual inspection. Indeed, there is no fixed preprocessing pipeline applicable to all the dataset, this is something still lacking also if there is some new proposed solution based on the use of deep learning algorithm.

After the preprocessing, time series can be analysed at the sensor level extracting the time-frequency map, but this is not so much informative, because on the contrary of LFPs, here the sensors and the sources of the signal are not in the same location. Each recorder time series coming from the sensors (outside the brain) contain signals coming from several sources (inside the brain); thus we are interested in extracting the signal at the local source level before proceeding with analysis. To do so, an MRI of the subject and a series of computational passages are needed.

The MRI is used to reconstruct a complete 3D model of the brain of the subject, including the skull and skin, through the use of softwares such as FreeSurfer or BrainVISA. The brain is further segmented between white and gray matter and then with parcelization an atlas is applied in order to label different cortical or subcortical

regions. Atlases can follow an anatomical or a functional division, following different subdivision rules, some example are the Brodman atlas following cytoarchitecture, the Desikan-Killiany (Desikan et al., 2006) following gyruses, and MarsAtlas (Auzias et al., 2016) following sulci.

Once we have the model of the anatomy, we need to model what we can observe from our sensors given the anatomical constraints that we just computed, in other words we need a forward model. To do so we must compute two things: a source space and a volume conduction model. The source space is needed to describe sources' position relative to each sensor and their orientation in space, that is the orientation of the electrical dipoles. Sources can be placed on a surface mesh, with orientation corresponding to the normal direction of the surface, or in a volumetric space, with free orientation. Volume conduction models are needed in EEG because the electric field can diffuse differently through brain, skull and skin, causing distortions in the recorded electrical signal, while in MEG is needed because this diffusing electrical field generates itself a small magnetic field that can distort the magnetic signal, but generally MEG is less affected by this phenomena because magnetic fields penetrates non-magnetisable materials. One of the most used volume conduction models is obtained with the boundary elements method (BEM) because it is easy to compute, since it consists of a mesh of triangles describing the surface of the skull and the skin surrounding the brain. Other available methods are the finite element method (FEM) and the finite difference method (FDM), which return 3D conductivity models.

Once we have the forward model, we must compute the contribution of each source given the recorded signal from the sensor, or in other words the inverse model. Several techniques can be used for inverse modelling, such as single/multiple dipole fitting (minimizing the error between model and measured field), distributed source models, and state of the art spatial filtering methods (also called beamforming), like the dynamic inherent court of sources (DICS), the Linearly Constrained

Minimum-Variance (LCMV) or the Synthetic Aperture Magnetometry (SAM). Those algorithms give as result a source-level time-resolved signal.

Usually, the number of computed sources are higher than the number of original sensors (e.g. in study 3 we computed 4000 sources for each brain hemisphere starting from a total of 248 sensors), thus to simplify computations is better to group and merge them (e.g. averaging) accordingly to the pre-computed parcelization, to reduce the number of signal dimensionality to the number of parcels (e.g. in study 3 we used MarsAtlas that has 48 parcels per hemisphere).

After all these passages, we can use the signal at the parcel level to compute event related potentials (ERP, also called evoked responses), or a time-frequency map of the power of several frequency bands. Extracting the power of a signal is very common because it allows to study the rhythmic oscillatory activity of well established frequency ranges: delta (1–4 Hz) linked to sleep state, theta (4–8 Hz) linked to drowsiness, alpha (8–12 Hz) linked to resting state, beta (15–30 Hz) linked to attention, gamma (30–80 Hz) linked to focus, and high-gamma (>50 Hz) linked to problem solving and concentration (Cole and Voytek, 2017). This is just an overview, but these bands are shown to correlate with precise motor responses (Jenkinson and Brown, 2011; Schwerdt et al., 2020), cognitive states (Brovelli et al., 2005), behavior (Engel and Fries, 2010) and also with pathological states (Holt et al., 2019). Moreover, different frequency bands are associated with different ranges of cortical interactions, with the lower bands implied in large scale computations and higher frequency associated with local activations (von Stein and Sarnthein, 2000).

There are several different algorithms to extract the periodic component of a signal, as for example the Fast Fourier Transform (FFT), the Morlet Wavelet convolution (MW), and the Multitaper method (MTM), that are among the most used algorithms. The FFT took over the Discrete Fourier Transform (DFT) for its computational speed, especially when considering long time series, and it's still used for spectral analysis and denoising. The MW method allows fast resolving of the periodic components in both time and frequency domain. This is possible through the computation of

several wavelets, small frequency-specific waves with particular properties, that are convolutionally multiplied to the signal. The MW has one particular parameter used for the construction of the wavelets, that can bias the result of the analysis: the number of cycles. This is a well known issue of this method, responsible for what is called the temporal-spectral tradeoff, by which wavelets with a lower number of cycles give a better representation in the time domain, while wavelets with a higher number of cycles give a better representation in the spectral domain. In order to find a good compromise between time and spectral precision it is usually used a variable number of cycles, increasing together with the frequencies.

The MTM uses Slepian tapers sequence, small snippets of data of which the first one is a gaussian and all the others are orthogonal among them. The data are convolutionally multiplied to all of these tapers, highlighting different properties of the signal, and then a FFT is computed on each data-taper to obtain the spectral analysis. The sum of these spectra gives the power of the data for each convolution.

In the end, we can understand why these techniques are largely used to study brain computations and brain dynamics, indeed they have an outstanding temporal resolution (in the order of milliseconds, with a final sampling frequency of around 500-1000 Hz), they are non invasive and they allow recording the whole brain activity, but a little clarification here is needed. When we find the word 'brain' associated with these techniques, it is more convenient to read it as 'cortical'. Indeed, when we build the source space, we can consider both cortical and subcortical sources, but the more they are distant from the recording zone the more the modeled signals can incur in artifacts. That can happen for several reasons, such as leaking activity or error in the volume conduction model, and in the specific case of MEG we should also consider that the force of the magnetic field is inversely proportional to the square of the distance from the source of the field. Despite that, new solutions, algorithms and procedures to enhance the signal reconstruction at the level of deep sources are often proposed (Pizzo et al., 2019; Seeber et al., 2019), making it a still active research field.

In **Section 4**, I investigated the large-scale correlates of goal-directed learning using MEG recorded on human participants while performing a goal-directed causal-learning task. After data acquisition I computed the high-gamma activity (HGA) at the single-trial level and used information theory tools to relate it to behavioral variables.

1.3.3 Complementary techniques

As I told before, there are numerous techniques that are used for neurophysiological recordings and among them fMRI is one of the most used. fMRI is a non-invasive technique that uses a very powerful electromagnet to orient in space hydrogens' nuclei of water molecules, in order to let them produce a detectable magnetic field, with different variations in strength which allow us to distinguish different structures. To give an idea of how powerful these machines are, the earth's magnetic field is about 30 to 60 microtesla (μT), while the highest resolution fMRI machine so far can produce up to 7T, giving us the opportunity to produce images in wich we can discriminate cortical layers. Moreover, thanks to the properties of hemoglobin, fMRI can detect variations in blood oxygenation level (blood oxygenation level dependent signal, or BOLD signal) that correlates positively with brain's areas activity. Unfortunately, fMRI falls in the category of good spatial but bad temporal resolution, indeed their sampling rate is about 0.5 Hz (one point each two seconds), making them not particularly suitable for network dynamics studies. Anyway, new hybrid techniques that allow EEG recording during fMRI acquisition are so promising for solving problems of both the techniques.

Another exploited technique, especially in last years, is two-photons calcium imaging, whit which is possible to record the activity of populations of neurons, with a single neuron resolution, in behaving subjects. It is a very invasive technique, the region of interest is injected with a calcium-sensitive dye or more often neurons are genetically modified to express a calcium indicator, in order to emit a fluorescent

signal that reflects the spiking activity. A window is opened on subjects' skulls in order to access them with a two-photons microscope able to capture the intensity of fluorescence emitted by neurons, with a temporal resolution of around 10-30 Hz.
Ideally, the best solution would be to develop the perfect recording technique that allows us to acquire large brain areas activity at the neural level and with a few thousands of hertz of temporal resolution in a non-invasive way. But until that moment, the best practice is to choose wisely the technique that we want to use according to our study and the phenomena that we want to observe. Indeed we can't say that one of these techniques is better than the other, but just that one can be more suitable than the others in that specific context.

1.4 Computational models of goal-directed learning

Nowadays, computational models are used in several fields of research, not only regarding life science. Computational models can be useful to explain observed phenomena, to make predictions, to formulate new theories, to test hypotheses and to find analogies with reality. In the context of neurosciences, the aim of computational modelling is to provide common theoretical ground for disparate neurophysiological studies. Cognitive neurosciences and computational models can be considered as two sides of the same coin. Indeed most of the cognitive theories of behavior, referred both to classical and instrumental learning, find their roots in computational models based on behavioral studies (Rescorla, 1966; Rescorla and Wagner, 1972; Allan and Jenkins, 1980; Watkins and Dayan, 1992; Dayan et al., 1995; Sutton and Barto, 1998). An important part of cognitive modelling is the choice of the model to use. This depends on what we are trying to model, on what we expect as output of the model and how we want to use this output. In this work we made extensive use of computational models: in **Section 2** we used a spiking neural network model to explain how higher cognitive computations, such as planning and GDB, can emerge from neural processes (Basanisi et al., 2020). In

Section 3 we used a Q-learning model, a type of reinforcement learning model, fitted on monkeys' behavioral data in order to retrieve single-trials RPE values. In **Section 4** we implemented a Bayesian optimal agent model to compute relevant behavioral values, such as the contingency values, adapted on the behavior of human participants performing a goal-directed causal learning task.

1.4.1 Neural networks

A glimpse of history

Artificial neural networks (ANN) started their history in 1958 with Frank Rosenblatt's perceptron (Rosenblatt, 1958). The idea behind perceptron was easy: two or more input units are connected to one or more output units through weighted connections; the activation of each output unit depends on the sum of the weights of its active input units passed through an activation function (e.g. a step function or a sigmoid function). This is the general principle that most neural networks follow. But this simple perceptron was only able to solve linear problems (e.g. the OR and the AND problems) but not nonlinear problems (e.g. the exclusive-or, or XOR problem). This was possible after a while (after the so called 'AI winter') with the advent of multilayer perceptron, that showed that adding one or more middle 'hidden' layer between the input and the output, and using a supervised learning algorithm called backpropagation, was sufficient to solve most of classification problems, if we only have enough training time, enough layers, and enough weights to train. Soon after this problem was solved, and with the advent of improved computational power, newer ANN models exploded in a variety of novel structures and learning rules (Shrestha and Mahmood, 2019).

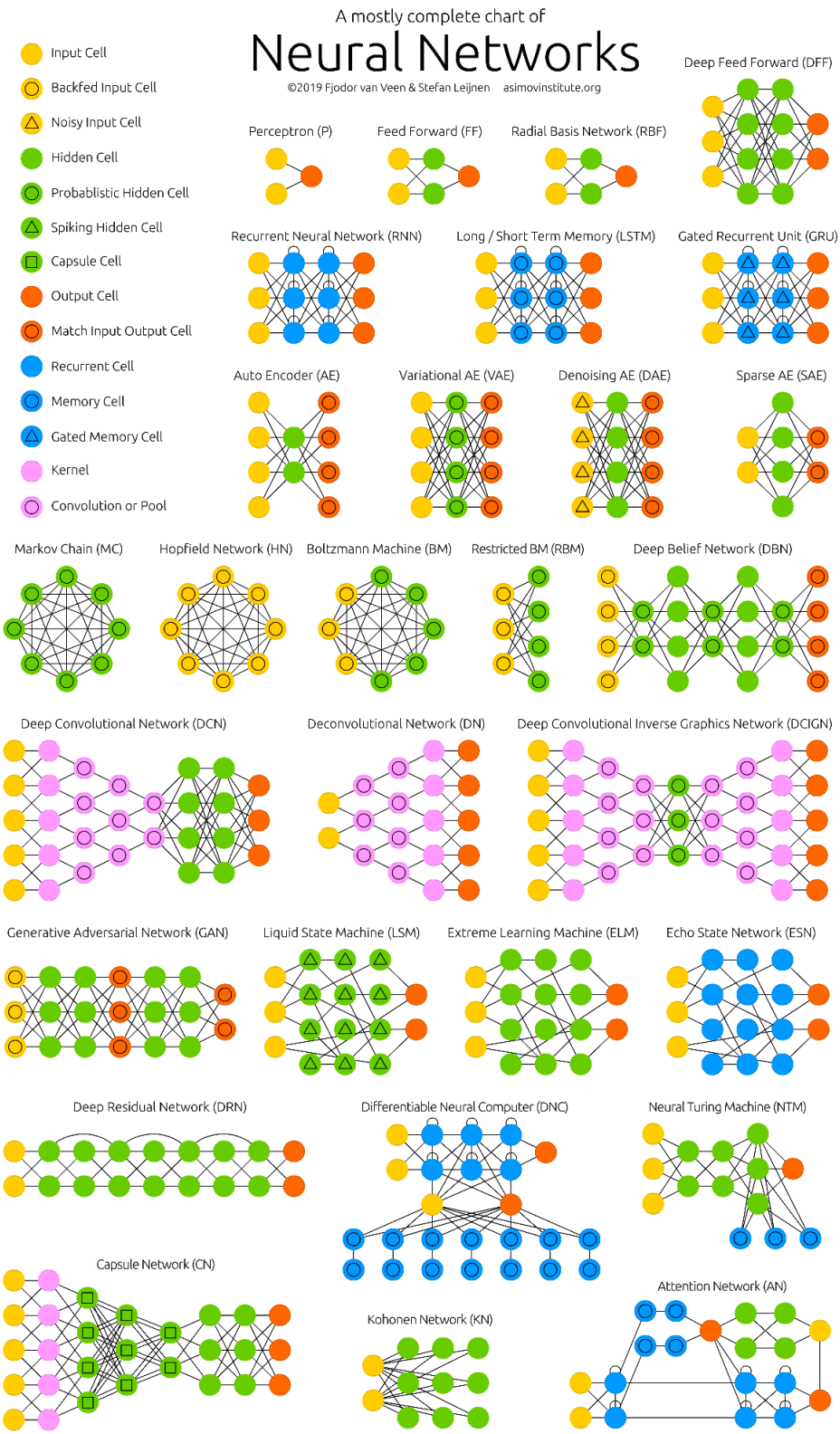


Figure 6. *A mostly complete chart of neural networks. Image taken from the Asimov Institute (https://www.asimovinstitute.org/neural-network-zoo/).*

Spiking Neural Networks

The need of building biologically inspired ANN led to consider time as an important feature of the network, thus new structures such as echo-state networks made of leaky neurons, or the spiking neural networks (SNN) made of integrate-and-fire neurons or spiking neurons ariesed (Maass, 1997; Ghosh-Dastidar and Adeli, 2009; Ponulak and Kasinski, 2011). The SNN that we describe in **Section 2** provides one example: at each discrete instant of time, once computed the sum of the weights' contributions that each unit receives, the network stochastically selects one firing unit through a SoftMax function. The firing event triggers the update of the weights following the Hebbian rule based on the spike-timing dependent plasticity (STDP), that increments the weight between that unit and the previously spiking one, and lowers the strength towards the units that fired distant in time. This process, combined with the network architecture, allows some kind of lateral inhibition that installs a 'winner-take-all' (WTA) mechanism, making the network able to learn in an unsupervised fashion the transitions between states as a Hidden Markov Model (HMM). Those powerful models are still currently studied because they represent a good possible bridge (and compromise) between neural models, bayesian computations, and biological complexity.

1.4.2 Reinforcement learning models (RLM)

Reinforcement learning (RL) was influenced by behavioral psychology and modern neuroscience, and it was developed as an emerging field of artificial intelligence and machine learning (Sutton and Barto, 1998). RL can be considered as a particular case of unsupervised learning based on the interaction with the environment. Indeed, contrary to supervised learning, RLMs do not need an explicit input nor an outcome to tend toward. They rely on a reward and a RPE signal to learn by trial and error actions' consequences on the environment. The general principles of RLMs can be summarized as follows: given an agent, able to perform a set of actions 'A' in an

environment that can be discretized in a set of states ‘S’, it will learn to predict the actions dependent state-state transitions in order to maximize the received reward ‘r’. This process resembles what is called a Markov Decision Process (MDP), where the choice of the action to perform, to reach the next desired state, is based only on the last observed state. The rise of these models started with the Temporal Difference (TD) learning model (Sutton and Barto, 1998), directly deriving from the Rescorla-Wagner model for classical conditioning (Rescorla and Wagner, 1972). Those two models introduced in their algorithm the concept of ‘error based learning’, that became so popular especially after the discovery that dopaminergic midbrain neurons activity correlates with error signals (Schultz et al., 1997). TD-learning can efficiently solve the prediction problem, indeed it is able to learn to predict the states associated values over multiple time steps. However, the control problem, i.e. to make an agent able not only to learn to predict the states values but also to use this prediction to orient its actions in order to maximise the reward, was still unsolved. As an extension of the TD-learning, addressing the problem relative to the choice of the actions distinctive of instrumental learning, in 1989 Christopher J.C.H. Watkins introduced the Q-learning model, then formalised in 1992 by Watkins and Peter Dayan (Watkins and Dayan, 1992). Briefly, Q-learning is a model-free algorithm able to numerically describe relations between state-action couples, assigning and updating these values depending on the RPE. The RPE is computed as the difference between the received reward and the expected reward. The action to perform is computed with a SoftMax function, that preferentially selects the action that will lead the agent toward the state with the higher expected reward. In **Section 3**, we used a Q-learning model fitted on monkeys behavior to estimate the RPE values from behavioral choices. Such learning signals were then correlated with LFPs data recorded in striatum, finding a RPE responsive beta-oscillatory activity establishing a gradient from the most rostro-ventral striatal part to its most caudo-dorsal part. For their versatility, and for the fact that they can have both an

algorithmic and a neural implementation, RLM are currently widely used in several fields of research.

The RLM that I described here is considered model-free, indeed, although it might sample from experience memory, it relies only on on-line samples from the environment. This means that it doesn't generate predictions of the next state and next reward to drive behaviour. Thus, it is particularly suitable for modelling HB but not for GDB, that is model-based, and needs an exhaustive model of stimuli-actions-outcomes to implement specific functions like planning.

HB-GDB modulation: the arbitrator model

One of the open issues in decision-making is how an agent is able to switch between habitual and goal-directed behavior, and thus between a model-free and a model-based strategy, and vice-versa. The generally accepted idea is that an agent starts to explore the environment in a goal directed way building a model of action-outcome associations. As learning goes by, if the environment is stable , the agent progressively consolidates those associations into habitual responses, becoming outcome insensitive. Thus, the more the agent repeats these action-outcome associations, the more they will shift toward a stimulus-response association. Later in time, if the known stimulus appears, an arbitration mechanism will trigger an habitual response. That's also the reason why HB is outcome insensitive and it will be way harder to shift back from HB to GDB. As we can see, this hypothesis is based on a main assumption: HB and GDB relies on two different competing systems, and thus presumably on different brain networks (Daw et al., 2005; Brovelli et al., 2008; Lee et al., 2014). Therefore, as we gradually pass from a HB to a GDB, we should be able to observe a gradual switch between the use of the two networks. Some studies in rodents (Yin and Knowlton, 2006; Hilario, 2008) actually confirmed this hypothesis, suggesting the involvement of the striatum in the arbitration mechanism, by observing a spatial shift in activation from its most

dorso-lateral part to its most ventro-medial part, and finding some similarity in humans (Balleine and O'Doherty, 2010).

The discussion is still open on how this arbitration mechanism orchestrates HB and GDB to efficiently switch from one to the other when both learning and action execution are needed. One of the most accepted models hypothesizes the existence of a flat arbitration mechanism that acts like a switch between the two comportamental strategies. Thus, when an agent is introduced to a new task, it is supposed to start using the surrounding stimuli to try to trigger a fast habitual response or a reflexive GDB (Keramati et al., 2011). Here, if the agent has no previous knowledge about stimuli-actions associations, the arbitration system allows the agent to inhibit the habitual system in order to switch toward a goal-directed strategy. Thus the agent starts exploring all the possible actions and to observe the consequent outcomes. Once he finds out which action leads to the desired outcome in response to the stimuli, he will start exploiting that action. Thus, a first phase of the GDB is exploration, during which the agent starts to perform random actions and to observe the resulting outcomes to collect knowledge about the structure of the task. After he obtained an undesired outcome, an agent can decide if to continue exploration, or on the contrary, after obtaining a desired outcome, he can pass to the second phase of GDB that is exploitation, that is the repetition of actions that led the agent in the desired state (Mehlhorn et al., 2015; Domenech et al., 2020). Exploration and exploitation are two swappable phases of GDB, indeed if we introduce a volatility in the task that changes the associations between actions and outcomes, the agent will restart exploring the environment in order to change the previously learned model of the world, and coming back exploiting the correct motor response in a few trials. A computational study based on the combination of a Q-learning model with a Bayesian working memory seems to accreditate this cognitive model by reproducing behavioral performances and reaction times of human participants performing a visuomotor learning task (Viejo et al., 2015).

Other studies stated instead the existence of a hierarchical control of the GDB, where the transition from goal-directed to habitual actions relies mostly on a process similar to the motor chunking of movement primitives (Ostlund et al., 2009; Botvinick et al., 2009; Dezfouli and Balleine, 2013; Balleine et al., 2015). According to this cognitive model, a global goal directed system is always active, and it evaluates at each decision if there is an HB that can be triggered in order to efficiently actuate a motor response in order to achieve the goal. If a habit is selected, after the action or sequence of action are executed, the behavior returns to be goal directed (Dezfouli and Balleine, 2013).

1.4.3 Bayesian models for goal-directed causal learning

Bayesian statistics finds its roots back in 1763, with an essay written by Reverend Thomas Bayes. This theorem outlines how to determine the probability of future events by taking into account how past events are distributed, also called inverse probability. Although this theorem was formalised about 250 years ago, its use increased exponentially in the past few years in several research fields, from statistics to modelling. Bayesian models took over cognitive science together with the idea that the brain operates like a probabilistic Bayesian machine, able to represent uncertainties of the world in terms of probability distributions and inferential processes based on Bayes' rule (Dayan et al., 1995, 2007). This concept was then extended to model higher cognitive processes and decision making (Baker et al., 2006; Griffiths et al., 2008). The elements composing Bayes theorem are three: 1) a prior probability, expressing beliefs and uncertainty about the distribution of past data, before any evidence is taken into account; 2) a likelihood function, that is a model of the relations between the prior and the posterior probabilities based on the observable data; 3) a posterior probability, that describes the distribution of the data taking in consideration prior knowledge and the likelihood function. In **study 4** we used a Bayesian ideal observer model, able to reproduce the behavior of human

participants performing a goal-directed causal learning task. We decided to use a Bayesian model because it allows us to model step by step the exploration phase and the progression of participants' learning, since the task we used was not focused on maximising expected reward (as in RL), but in maximising the knowledge about the causal relation between actions and their outcomes. Thanks to this model, we were able to estimate, at the single-trial level, the evolution of relevant task-related behavioral variables that we used to assess the role of different brain regions. Bayesian models are very powerful to describe the behavior of an agent at an high level, and for this reason their implication in artificial intelligence is increasing, however they are poorly informative about the low level neural computations. Studies attempting to bridge Bayesian computations and neural architectures are currently emerging.

1.5 Identifying the neural correlates of goal-directed learning

The aim of the studies described in **Section 3** and **Section 4** was to identify the neurophysiological correlates of goal directed learning. To do so, we correlated the single-trial brain dynamic aligned on a relevant event with the single-trial estimation of both task-related and modelled behavioral data. With a single-trial level analysis, we can describe how neural signals and behavior coevolve in time, and this is suitable for the study of cumulative processes, such as learning. To perform single-trial analysis and quantify the relation between neural and behavioral variables, we used an information theoretic approach and measures, such as the Mutual Information (MI), able to quantify the statistical dependence between two variables. Finally we used group-level inference and cluster based statistics to assess for significance.

1.5.1 Model-free and model-based analysis of brain data

Often, in neurophysiology, once defined the objective of the study, the experimenter is called to design a task in order to make some behavioral difference explicit during the task execution. Indeed, the final interest of analysing a set of neurophysiological data is to find differences among the data, or to couple them with the behavior in order to find some correlations that can explain the recorded data. There are two common methods to find those differences: following a model-free approach or following a model-based approach. Both of these approaches have pros and cons.

In the model-free approach, the data are divided by (or compared with) explicit task or subject dependent variables. They can correspond to, for example, reward and punishment values, reaction times, or different imposed conditions. The advantage of using this approach is that we are working with empirically observable and measurable variables, which allows us to explore data without making any assumption on their distribution. The limitations of model-free analysis are linked to task-design or recording machines, indeed it is often difficult, if not impossible, to retrieve trial by trial specific behavioral measures. Moreover, some behavioral variables originate from behavioral models and thus it is not possible to directly measure them.

On the contrary, using a model-based approach, a technique that derives from the fMRI literature (O'Doherty et al., 2007; Brovelli et al., 2008), means making assumptions on how the data are distributed, and that can be done in two ways: 1) using statistical (e.g. decoders or regression) data-driven models, fitted directly on the data to learn how they are distributed. In this case the assumption depends on the choice of the statistical model (e.g. linear vs. non-linear regression); 2) using behavioral models to compute implicit non-observable variables, in this case the assumption depends on the chosen algorithm to compute that variable, that can change the shape of its distribution. Usually, this latter approach requires the considered model to be previously validated or fitted on behavioral data, giving us the advantage to test hypotheses on modeled behavioral variables, often computed

at a finer time scale than the recorded ones. This second approach is very powerful for relating the subject's behavior to neural correlates and potentially disentangling subtle cognitive processes (like contextual learning). However, the limitations of model-based analysis are linked not only to the choice of the model, indeed sometimes modeled behavioral variables can incur in misinterpretation or overinterpretation, as they derive from definitions borrowed by cognitive science. For the studies described in **Section 3** and **Section 4** we performed model-free analysis based on information theoretic measures, namely the mutual information, of which I will give an overview in the next paragraph.

1.5.2 Information theory

Thanks to the pioneristic work of Claude Elwood Shannon on information theory (Shannon, 1948), we were able to build a mathematical framework that links the probability of an event to its uncertainty, or entropy, and consequently to its information. Information theory, formerly used mainly in communication, is now used in countless fields, and it is nowadays of common application in neuroscience (Timme and Lapish, 2018). Given an event with probability p_i to occur, we can write the associated shannon entropy (H) as:

$$H = -\sum_i p_i log_2(p_i) \tag{1}$$

The base of the logarithm defines the unit of the entropy. As we used a base 2 logarithm, the result of this equation will be expressed in bit (binary digit), another common base for the logarithm in **Equation 1** is the Euler number e, in that case the entropy is expressed in nat (natural unit of information). Importantly, Shannon linked the concept of entropy and uncertainty to the concept of information and surprise. The information content of an event quantifies how surprising that event is on average, thus the more an event is uncertain, the more its occurrence is surprising and yields information. It means that, on the contrary, a deterministic

event yields no information at all. Moreover, the less probable an event is, the more information it yields. In the case that more independent events are measured separately, the sum of the information content over single events gives us the total amount of information. To give an example that resumes what was said until now, I'll take in consideration the case of a coin toss and a dice roll. Imagine we should say how much information is carried by a coin toss: the possible results are head or tail, thus the probability to obtain one of the results is $p = 0.5$. Given that entropy tells us the average information in a probability distribution over the sample space, we can write:

$$H = -\left(\frac{1}{2}log_2\left(\frac{1}{2}\right) + \frac{1}{2}log_2\left(\frac{1}{2}\right)\right) = 1\ bit$$

Knowing the result of a coin toss give us 1 bit of information; instead, in the case of a dice roll, where the probability to have 1 of the six numbers is $p = 1/6$, the information obtained from knowing the result of the roll is:

$$H = -\ 6\ *\ \left(\frac{1}{6}log_2\left(\frac{1}{6}\right)\right) \simeq 2.58\ bit$$

As we can see, knowing the result of a dice roll gives us more information than the coin flip. That can be interpreted as the fact that knowing the result of a dice roll means also knowing that five other equiprobable results were discarded.

If we have a set of events $X = \{x_1, x_2, ..., x_n\}$ we can write:

$$H(X) = -\sum_{x \in X} p(x) log_2(p(x)) \tag{2}$$

Following the Venn diagram in **Figure 7**, if we consider two sets of independent events X and Y, we can define the joint entropy between X and Y as the sum of the individual entropies:

$$H(X, Y) = -\sum_{xy} p(x, y) log_2(p(x, y)) \tag{3}$$

The conditional entropy of X given Y, is the average conditional probability over Y:

$$H(X|Y) = H(X, Y) - H(Y) \tag{4}$$

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) log_2(p(x|y))$$

$$= -\sum_{xy} p(x,y) log_2(p(x|y)) \quad (5)$$

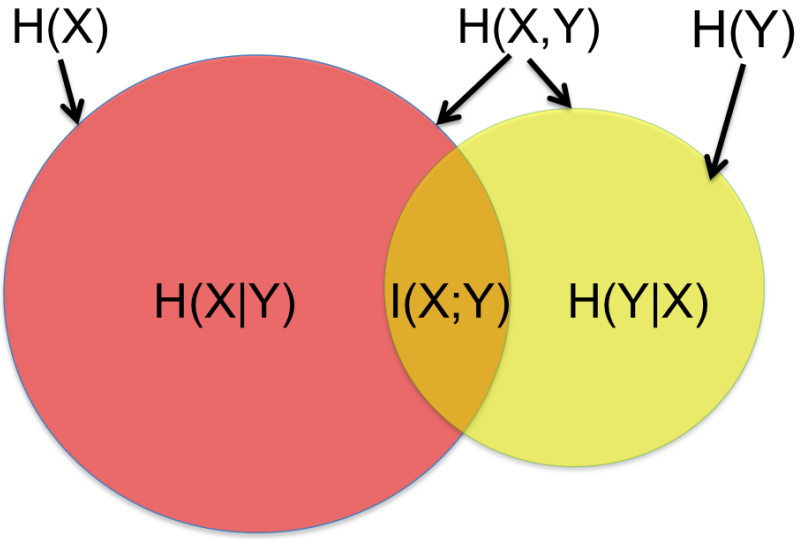


Figure 7. Venn diagram of Shannon entropy and mutual information. Image taken from Google.

Mutual Information (MI)

The mutual information $I(X;Y)$ quantifies the statistical dependency between two variables X and Y, expressing it as the amount of information carried by one of the two variables when we observe the other one. MI is a non-negative and symmetric ($I(X;Y) = I(Y;X)$) measure that is formulated by the definition of conditional and joint entropy:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X|Y) \end{aligned} \quad (6)$$

In the context of this thesis we used MI as a descriptive measure of the statistical dependence between neurophysiological signals and model-free or model-based behavioral variables. A standard approach for estimating MI between two continuous variables implies a binning step in order to estimate the full joint probability distribution (Timme and Lapish, 2018). However, a consequent amount of data, hard to reach in the context of brain signals, is usually required in order to

have a decent sampling of this probability distribution. To overcome those inherent limitations, we used a binning-free alternative, originated from the field of economics and recently ported to neuroscience, called Gaussian Copula Mutual Information (GCMI) (Ince et al., 2017). In short, the GCMI exploits the fact that the MI does not depend on the marginal distributions of the variables but only on the copula function which describes their statistical dependency. GCMI has shown to be a robust alternative to MI, and to capture both linear and non-linear statistical dependencies as long as this relation is roughly monotonic.

1.6 Thesis objectives

The objective of this thesis is to give a contribution to the investigation of the computational and neurophysiological correlates of goal-directed learning and behavior through the analysis of neural data and the use of neural and behavioral models. To achieve that, we first built a spiking neural network model to provide a plausible explanation of how goal-directed model-based learning can emerge from neural computation in an unsupervised fashion (**Section 2**). Thanks to its architecture and to the STDP-based learning rule the model is able to encode sequences of stimuli-actions-outcomes and to use them according to the goal to orient behavior and make predictions. Then I investigated the role of beta-band oscillations in non-human primates striatum in encoding RPEs signals computed with a Q-learning model, associated with a free-choice probabilistic learning task (**Section 3**). We found that information about RPEs are distributed across striatal fields forming a gradient stronger toward the rostro-ventral part and weaker toward the caudo-dorsal part. Finally, we investigated the temporal dynamic of different cortical brain regions of human participants performing a goal-directed causal learning task, in encoding relevant cognitive measures computed through the use of an optimal observer Bayesian model (**Section 4**). We characterized both action and

outcome-related activation of mostly orbitofrontal and prefrontal regions, but also parietal and temporal regions, significantly responding to ΔP, P(O|A) and P(O|C).

Section 5 of this manuscript will give an overview of my scientific contributions in and outside the BraiNets team and of my personal scientific interests, especially those related to the Open Science movement and neuroinformatics projects.

1.7 Publications

- **Basanisi, R.**, Brovelli, A., Cartoni, E., & Baldassarre, G. (2020). A generative spiking neural-network model of goal-directed behaviour and one-step planning. *PLoS Computational Biology*, *16*(12), e1007579.
- **Basanisi, R.**, Marche, K., Combrisson, E., Apicella, P., & Brovelli, A. (2021). Beta oscillations in the monkey striatum encodes reward prediction error. (*In preparation*)
- **Basanisi, R.**, Combrisson, E., Dauce, E., & Brovelli, A. (2021). Dynamics of human cortical circuits mediating goal-directed causal learning. (*In preparation*
- Gau, R., Noble, S., …, **Basanisi, R.**, …, Marinazzo, D. (2021). Brainhack: Developing a culture of open, inclusive, community-driven neuroscience. *Neuron*, *109*(11), 1769-1775.
- Combrisson, E., Allegra, M., **Basanisi, R.**, Ince, R. A., Giordano, B., Bastin, J., & Brovelli, A. (2021). Group-level inference of information-based measures for the analyses of cognitive brain networks from neurophysiological data. *bioRxiv*. (Under revision, NeuroImage)
- Combrisson, E., **Basanisi, R.**, Cordeiro, V.L., Ince, R. A., & Brovelli, A. (2021). Frites: A Python package for functional connectivity analysis and group-level statistics of neurophysiological data. (Under revision, JOSS)

Section 2. A generative spiking neural-network model of goal-directed behaviour and one-step planning

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A generative spiking neural-network model of goal-directed behaviour and one-step planning

Ruggero Basanisi[1], **Andrea Brovelli**[1], **Emilio Cartoni**[2], **Gianluca Baldassarre**[2]*

1 Institut de Neurosciences de la Timone UMR 7289, Aix Marseille Université, CNRS, Marseille, France, **2** Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy

* gianluca.baldassarre@istc.cnr.it

Citation: Basanisi R, Brovelli A, Cartoni E, Baldassarre G (2020) A generative spiking neural-network model of goal-directed behaviour and one-step planning. PLoS Comput Biol 16(12): e1007579. https://doi.org/10.1371/journal.pcbi.1007579

Editor: Boris S. Gutkin, École Normale Supérieure, College de France, CNRS, FRANCE

Received: December 3, 2019

Accepted: October 1, 2020

Published: December 8, 2020

Data Availability Statement: This work is based on a computational model. The code of the model has been made publicly available for download in a GitHub repository: https://github.com/GOAL-Robots/ASpikingModelOfGoalDirectedBehaviour.

Funding: R.B. carried out this work thanks to the support of the A*MIDEX grant (n°ANR-11-IDEX-0001-02) funded by the French Government «Investissements d'Avenir » program (https://anr.fr/en/investments-for-the-future/investments-for-the-future/). A.B. was supported by the French

Abstract

In mammals, goal-directed and planning processes support flexible behaviour used to face new situations that cannot be tackled through more efficient but rigid habitual behaviours. Within the Bayesian modelling approach of brain and behaviour, models have been proposed to perform planning as probabilistic inference but this approach encounters a crucial problem: explaining how such inference might be implemented in brain spiking networks. Recently, the literature has proposed some models that face this problem through recurrent spiking neural networks able to internally simulate state trajectories, the core function at the basis of planning. However, the proposed models have relevant limitations that make them biologically implausible, namely their world model is trained 'off-line' before solving the target tasks, and they are trained with supervised learning procedures that are biologically and ecologically not plausible. Here we propose two novel hypotheses on how brain might overcome these problems, and operationalise them in a novel architecture pivoting on a spiking recurrent neural network. The first hypothesis allows the architecture to learn the world model in parallel with its use for planning: to this purpose, a new arbitration mechanism decides when to explore, for learning the world model, or when to exploit it, for planning, based on the entropy of the world model itself. The second hypothesis allows the architecture to use an unsupervised learning process to learn the world model by observing the effects of actions. The architecture is validated by reproducing and accounting for the learning profiles and reaction times of human participants learning to solve a visuomotor learning task that is new for them. Overall, the architecture represents the first instance of a model bridging probabilistic planning and spiking-processes that has a degree of autonomy analogous to the one of real organisms.

Author summary

Goal-directed behaviour relies on brain processes supporting planning of actions based on their expected consequences before performing them in the environment. An important computational modelling approach proposes that the brain performs goal-directed

National Agency (n° ANR-18-CE28-0016-01) (https://anr.fr/Projet-ANR-18-CE28-0016;https://anr.fr/Project-ANR-17-HBPR-0001). E.C. received a salary to carry out this work from the European Union, Horizon 2020 Research and Innovation Program, under Grant Agreement n° 713010 Project "GOAL-Robots - Goal-based Open-ended Autonomous Learning Robots" (https://cordis.europa.eu/project/rcn/203543/factsheet/en). G.B. received funding from the European Union, Horizon 2020 Research and Innovation Program, under Grant Agreement n° 713010 Project "GOAL-Robots - Goal-based Open-ended Autonomous Learning Robots" (https://cordis.europa.eu/project/rcn/203543/factsheet/en). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

processes on the basis of probability distributions and computations on them. A key challenge of this approach is to explain how these probabilistic processes can rely on the spiking processes of the brain. The literature has recently proposed some models that do so by 'thinking ahead' alternative possible action-outcomes based on low-level neuronal stochastic events. However, these models have a limited autonomy as they require to learn how the environment works ('world model') before solving the tasks, and use a biologically implausible learning process requiring an 'external teacher' to tell how their internal units should respond. Here we present a novel architecture proposing how organisms might overcome these challenging problems. First, the architecture can decide if exploring, to learn the world model, or planning, using such model, by evaluating how confident it is on the model knowledge. Second, the architecture can autonomously learn the world model based on experience. The architecture represents a first fully autonomous planning model relying on a spiking neural network.

Introduction

In mammals, the acquisition and consolidation of instrumental behaviour involves two sets of processes, one underlying flexible *goal-directed behaviour*, used in particular to find solutions to new problems or to face changing conditions, and the other one related to *habits*, forming stimulus-response behaviour used to efficiently, but inflexibly, face familiar conditions [1–3]. As also highlighted in the computational literature [4], goal-directed behaviour is *model based*; that is, it relies on an internal representation of the external world (in particular of the transition probabilities between its states; the so called *world model*) to internally simulate the consequences of actions, or action sequences, usable to achieve desired world states (*goals*) before executing them in the environment (*planning*) [5–10] (note that here goals are intended as internal representations of desired world states [11], rather than in the broader meaning of world/body states to which the organism homeostatically converges [12]). When the agent pursues a new goal and has a world model to do so, goal-directed processes allow the solution of the task on the basis of planning. Indeed, the world model represents the general goal-independent dynamics of the world, in particular how it responds to the agent's actions. The simulated achievement of the new goal based on the world model might be possibly marked by an internal reward [13] and to an external observer the agent appears to solve the new task 'on the fly' or 'by insight'. Instead, habitual behaviour is *model free*, in the sense that it relies on *actions directly triggered by stimuli* (*habits*) and does not require a world model anticipating their outcomes [4, 9, 14]. Habits are task dependent as they can only lead to specific desirable world states. Thus, given a new desired state, repeated experience is needed to discover and learn by trial-and-error the new stimulus-response associations leading to it.

In the brain, goal-directed behaviour relies on ventral/associative basal ganglia and prefrontal cortex areas supporting the representation of goals and the world dynamics; instead, habitual behaviour relies on motor basal ganglia and sensorimotor/premotor cortices able to acquire stimulus-response associations by reinforcement learning [14–18]. The brain processes underlying goal-directed behaviour have been interpreted within different computational frameworks. A current influential view of the brain, rooted in Helmholtz' pioneering contributions on perception [19], considers it a *probabilistic* or *Bayesian machine* that copes with the uncertainties of the world by representing it in terms of probability distributions and probability inferences on them pivoting on the Bayes rule [20, 21]. This view of the brain has been progressively extended to cover all aspects of cognition, from perception to action and

decision making (e.g., [22, 23]). In line with this view, it has been proposed that the brain also implements goal-directed behaviour and planning through probabilistic representations and inferences, and this has been shown through specific models developed within an approach called *planning as inference* (e.g., [24–26]). This approach uses world representations expressed as probability distributions and performs action selection based on a probability inference maximising the expectation of the desired world state (more details in Sec 1).

The models of planning as inference commonly use probability distributions that directly involve high-level aspects of cognition and behaviour, for example observations, world states, and actions; moreover, the inferences on these distributions are based on sophisticated mathematical manipulations of the parameters of the distributions, for example those based on *Hidden Markov Models* (HMMs), or on numerical approximations of them. This gives rise to a fundamental challenge for these models [21, 27–29]: *how can the probability distributions and inference processes supporting goal-directed processes be grounded on the low-level spiking events of neurons in the brain?*

An important possibility is that the needed probability distributions rely on the probability distributions of neuron spikes, sampled by the actual spikes; and that the connections between neural populations, undergoing experience-dependent plasticity, support the conditional probabilities underlying probabilistic inferences [25, 30–34]. In this respect, spikes can be seen as sampling probability distributions analogously to what happens in *particle filters* [35–37]. These are algorithms that use a set of values ('particles') to represent the distributions of stochastic processes such as HMMs (particle filters draw a set of random values –the 'particles'– for each probability distribution to represent, consider the dependencies between different distributions by 'propagating' the particles between them, and use value weights and re-sampling mechanisms to approximate complex distributions and take observations into account; [37]). In this respect, the model presented here relies on a general principle, also shared with previous models [38–40], termed here *emergent generativity*. We refer emergent generativity to the process for which the stochastic events of spiking neurons, happening at the micro/low level, are amplified by neural mechanisms to generate alternative cognitive contents, at the macro/high level, that support adaptive behaviour (e.g., alternative possible imagined percepts, believes, and courses of action). This concept is further discussed in Sec 3.2.

Although planning as inference was previously modeled with a firing-rate neural network [41], only recently recurrent spiking neural network models have been used to implement planning as inference [38, 39, 42]. These models, which are the state-of-the-art in the field, use recurrent neural networks to represent the world model. Here different groups of neurons represent different world states, for example different places in a navigation maze, and their lateral connections encode the possible transitions between states that the agent might cause with action. The spikes of the world model sample the prior probability of the state sequences followed by the agent if it explores the environment randomly, and of the rewards associated to the sequence (e.g., a reward of 1 when a target state is achieved). A second neural layer of spiking neurons that encodes the 'context', intended as the current and target states, sends all-to-all connections to the world model and can condition the probability distribution it expresses. The neural solution to the inference problem relies on the update of the connections linking the context to the world model so that the distance (Kullback-Leibler divergence) between the prior probability distribution of the state sequences converges to the desired posterior probability distribution maximising the reward. The actions needed to follow the state sequences sampled from the posterior distribution are inferred by inverse kinematics, either offline [38] or using a dedicated neural layer [39]. Another related model [40] reproduces goal-directed behaviour with an analogous recurrent spiking neural network. Here the actions that correspond to a decision-state are reciprocally linked by inhibitory connections to implement

decision making. For a given task, reward units 'inject' activation into terminal actions, and this activation diffuses backward towards the upstream actions to represent the anticipated value attributed to them. This value is then used for action selection.

These models represent an important first step in modelling how the brain might implement planning as inference, but much remains to be understood since planning in animals involves several interdependent complex processes such as the formation of goals, their motivational activation, the acquisition of world models, the formulation of plans at multiple levels of abstraction, the performance of actions, and the coordination of these different processes [43].

In this work we contribute to face these issues by tackling two important problems not solved by the state-of-the-art models considered above. The first problem is: *how can the brain acquire the world model while at the same time using it for planning?* The model-free literature on reinforcement learning [4] studies the important problem of the exploration-exploitation trade-off where an agent must decide whether to take random actions to explore the environment and learn the policies that lead to rewards, or to exploit those policies to maximize rewards. A problem less studied involves a situation where model-based/goal-directed agents have to face an analogous but different trade-off [44–46]. In particular, when these agents solve new tasks they have to decide if exploring to refine the world model, or if exploiting such model to plan and act. Here we consider the early phases of the solution of new tasks, involving either a new environment or a new goal, and hence focus on the latter type of exploration-exploitation trade-off. This problem has been recently faced in a principled way [46] within the probabilistic framework of active inference [22]. However, the proposed solution is applicable only to very simple scenarios where hidden-states are few and are given to the agent, rather than being autonomously acquired; moreover, and importantly for our objective, the solution has not been grounded on brain-like mechanisms. On the other side, current state-of-the-art models implementing planning as inference based on spiking networks either learn the world model before solving the target task [38, 39] or use a hardwired world model [40], and so they do not face the problem altogether. How the brain manages to learn and use the world model at the same time is hence a fully open problem.

The second problem we face here, not solved by the current planning-as-inference spiking models, is: *how could the brain learn the world model in an unsupervised fashion?* Currently there are no biologically acceptable solutions to this problem as the current state-of-the-art models either learn the world model off-line through supervised learning techniques [38, 39] or are given a hardwired model [40].

Here we propose a model architecture facing both problems limiting the current planning-as-inference spiking-network models. The architecture tackles the first problem by proposing a novel *arbitration mechanism* measuring the uncertainty of the world model on the basis of the entropy of the posterior probability distributions expressed by the neurons forming it (cf. [47, 48]). When this uncertainty is low, planning continues, otherwise exploration actions are performed. Recently, it has been shown that the contextual learning and use-for-planning of the world model encounter a difficult problem for which the world model can prematurely converge to sub-optimal solutions ('bad-bootstrapping' problem, [46]).

A second novelty of the architecture is the solution of the second problem. The solution is in turn based on three innovations. First, the integration of the unsupervised STDP learning rule proposed in [49] into the recurrent spiking neural-network world model. This allows the world model to learn at the same time the hidden causes of observations and the probabilistic time dependencies between them. This is a notable advancement in terms of biological plausibility with respect to current models using supervised learning mechanisms that directly activate the internal units to encode hidden causes [38–40]. This also represents a computational

advancement as the only recently proposed (non-spiking) probabilistic model tackling the model-based exploration/exploitation problem [46] assumes to know the hidden causes of observations. The second mechanism relies on the idea that the world model is a HMM that 'observes', learns, and predicts sequences of items formed not only by percepts but also actions. Actions are in particular 'observed' by the world model after being selected by planning or exploration processes. This idea was suggested by evidence indicating that various brain areas receive ('observe') both sensory and motor information, for example the parietal cortex [50, 51], the prefrontal cortex [43], and the hippocampus [52]. This, integrated with the third mechanism introduced below, allows the world model to autonomously select actions without the need of an auxiliary component selecting actions on the basis of state sequences (e.g., as in [38]). The third mechanism is based on the conditioning of the posterior probabilities of the world model on the pursued goal. This implies that with no goal conditioning the world model represents the prior probabilities of arbitrary state-action sequences, while when a goal is selected ('clumped') the model represents the posterior probabilities directly producing action-sequences leading to the goal.

The model was validated by testing it against the results reported in [15, 16] where human participants learn to solve a visuomotor learning task. In particular validation checked if the learning processes of the world model led to match human performance, and if the planning time spent by the arbitration mechanism reproduced the reaction times exhibited by human participants. The target experiment was also investigated with a model in [48]; however, this model did not aim to bridge planning as inference to spiking network mechanisms. To our knowledge, our model is the first of this type to be validated with specific detailed behavioural data.

The rest of the paper is organised as follows. Section 1 describes the model architecture and functioning and the visuomotor learning task used to validate it. Section 2 presents the results of the model tests, in particular by comparing the model performance and reaction times with those of human participants of the target visuomotor task, and by showing the mechanisms that might underlie such performance. Section 3 discusses such results in the light of the literature. Finally, Section 4 draws the conclusions. Particular attention has been paid to make the paper accessible to a wide interdisciplinary audience, as requested by the journal; this was also facilitated by leveraging the heterogeneous background of the authors.

1 Methods

This section first illustrates the task used to test the model [15, 16] and gives an overview of its architecture and functioning. Then it explains the HMMs relevant for this work, the spiking neural network equivalent to a HMM used to implement the world model of the architecture, the arbitration and exploration components of the architecture, and the procedure used to search its meta-parameters. The initial draft of this paper was published in [53].

1.1 Target experiment

In the task used to test the model [15, 16], human participants are supposed to discover the correct associations between three different stimuli and three possible motor responses chosen from five possible ones (Fig 1). During the experiment, three different colours are projected on a screen in a pseudo-randomised order, in particular through twenty triplets each involving the three colours (each triplet is thus formed by three 'trials'). After each colour perception, the participants have to respond by pressing one of five buttons of a keyboard with their right hand. Once this action is performed, a feedback on the screen informs the participants if the association between the colour and the performed action was correct or wrong. The goal of the

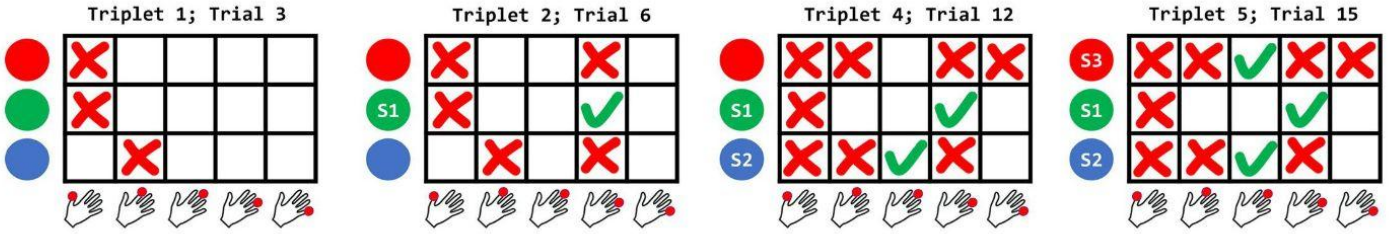


Fig 1. The visuomotor learning task used to validate the model. Three colour stimuli are presented to the participants in a pseudo-random order, in particular in triplets each containing each colour exactly once. The action consists in pressing one out of five possible buttons with the right hand. The figure refers to an ideal participant who never repeats an error for the same colour and always repeats the correct action after discovering it. The four pictures refer to respectively the actions after one, two, four, and five triplets: a red cross and a green tick-mark refer to incorrect and correct colour-action sequences respectively. The colour receiving the first action in the second triplet is marked as the 'first stimulus' (S1), and such action is considered the as correct one for it. The colour different from S1 receiving the first action in the fourth triplet is marked as the 'second stimulus' (S2), and such action is considered as the correct one for it. The colour different from S1 and S2 receiving the first action in the fifth triplet is marked as the 'third stimulus' (S3), and such action is considered the correct one for it.

https://doi.org/10.1371/journal.pcbi.1007579.g001

participants is to obtain a 'correct' feedback for each colour by selecting the corresponding 'correct action'. Unbeknown to the participants, however, the correct action for each colour is not set a-priori but is established dynamically during the experiment to obtain a fixed number of exploration actions for the three colours among the different participants. In particular, for each colour stimulus 'S' a fixed number of 'incorrect feedback' outcomes are given to the participants before considering the performed action as correct: thus, for S1 a 'correct' feedback is given at the second action (hence after one error), for S2 at the fourth action (after three errors), and for S3 at the fifth action (after four errors). The colour stimulus considered as S1, S2, and S3 is itself established dynamically as the first colour, not yet associated to a correct action, presented within respectively the second, fourth, and fifth triplet. Notice that with this procedure the participants are not supposed to explore all the possible colour-action associations but rather to only discover, and then exploit, the colour-action association needed to accomplish the 'correct feedback' goal. The task has been designed to differentiate between two phases of the participants' behaviour: an initial exploration phase where they are expected to search the correct colour-action associations, and a second exploitation phase where they are supposed to repeat the found correct associations until the end of the task.

1.2 Goal-directed behaviour model: Overview of the architecture and functioning

1.2.1 Architecture. Fig 2 gives an overview of the architecture and functioning of the model. The architecture of the model is composed of a spiking neural network for planning formed by four different layers, a spiking neural network for exploration formed by two neural layers, and a non-neural arbitration component. The four layers forming the core neural network, which supports planning by instantiating a HMM, are now considered more in detail.

Input layer. The input layer contains ten neurons, three encoding the stimuli (colours), five encoding the actions (different possible finger presses), and two encoding the outcome (correct or incorrect feedback). The input layer sends all-to-all afferent connections to the neurons of the associative layer.

Goal layer. The goal layer is composed of neurons encoding the goals to achieve, here two neurons encoding the two goals of the visuomotor task: 'obtain a correct feedback' and 'obtain an incorrect feedback' (the use of the latter is explained later). To commit to achieving a certain goal the agent activates the corresponding neuron on the basis of internal mechanisms not simulated here. Goal neurons send all-to-all efferent projections to the associative neurons.

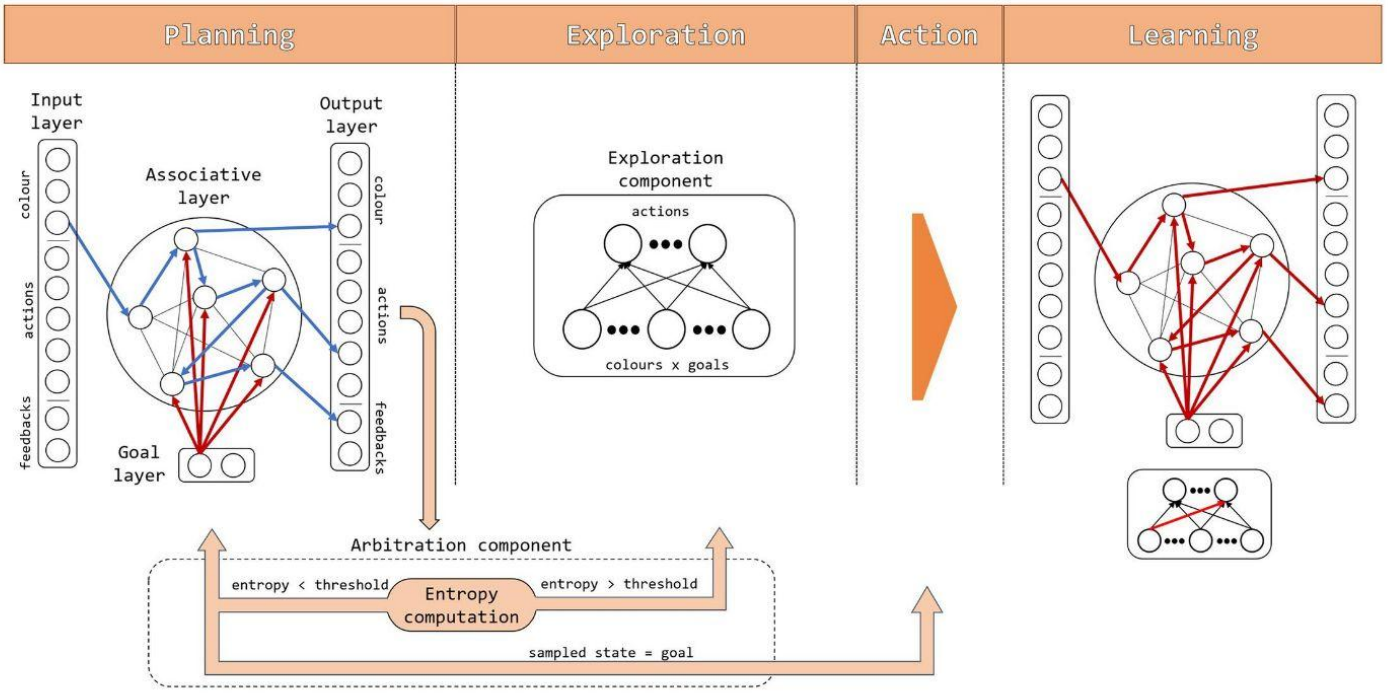


Fig 2. Architecture and functioning of the model: Components and information flow. The architecture is formed by a planning component (representing input patterns, hidden causes of input patterns within an associative layer, expected events including actions, and goals), an exploration component selecting actions when planning is uncertain, and an arbitration component deciding when to plan, explore, or act. The figure also shows the timing of the processes taking place during a trial, with the first two left graphs showing the Planning (exploitation) and (possibly) Exploration phases and the right two graphs showing the Action execution and Learning phases. Blue arrows represent an example of information flow travelling stable connections during the Planning phase and red arrows represent information flows travelling connections that are updated during the Learning phase.

https://doi.org/10.1371/journal.pcbi.1007579.g002

Associative layer. The associative layer, forming the core of the model, is composed of 400 neurons, all connected to each other but without self-connections. The associative layer receives the mentioned afferent connections from the input and goal layers, and sends all-to-all efferent connections to the neurons of the output layer.

Output layer. As the input layer, the output layer is composed of ten neurons each one representing one of the stimuli, actions, and outcomes of the task. The output layer receives the mentioned afferent connections from the associative layer.

Together the four layers instantiate a neural HMM implementing the world model used for planning. In particular, the input and output layers together form the observation part of the HMM, and have an identical structure. Given that the connections of real neurons are unidirectional, we used the two layers to implement separately the two functions played by the observation part of the HMM, namely the input from the external environment and the possible generative reconstruction of such input based on internal causes. The associative layer encodes the probability distribution over the hidden causes of the observations and the probabilistic temporal dependencies between them. The goal layer can condition the latter distributions to possibly increase, with learning, the probability of sampling simulated colour-action-feedback sequences that lead to the desired goal. An important feature of the HMM implemented by the model is that, as in [49], each of the three events of each trial (colour, action, feedback) is represented by a sequence of active HMM nodes that encode not only one of the events but also the time step when it is present. For example, after learning a certain group of

neurons encodes an action and the neurons of the group fire in sequence for a certain number of time steps corresponding to the action duration.

Alongside the planning components, the architecture is formed by the following additional components used for exploration and arbitration.

Exploration component. This component is formed by two layers of spiking neurons, namely (a) an input layer encoding the combinations of colours and goals (3×2 neurons corresponding to 3 colours and 2 goals), and (b) an output layer encoding the five possible finger-press actions (five neurons).

Arbitration component. This component, currently not implemented with neural mechanisms, decides when to plan, explore, or act in the world. The decision is made on the basis of the level of knowledge of the world model, measured as the average entropy of its probability distribution during the last 'planning cycle' (explained below). When entropy is lower than a certain threshold, and a goal has not been found, planning continues, whereas if a goal has been found the corresponding action is performed in the environment. If entropy is above the threshold then the control is passed to the exploration component that selects the action to perform in the world.

1.2.2 Functioning. The functioning of the model is summarised in Algorithm 1. The model experiences multiple trials of the task (lines 1-3 of the algorithm): 60 trials (20 colour triplets) with the goal set to 'achieve a correct feedback' (this reflects the target experiment [15]) and 60 trials (other 20 colour triplets) with the goal set to 'achieve an incorrect feedback' (as explained below, these additional trials are used to produce a prediction). Each trial of the task lasts for a certain number of discrete time steps (here 15). Each trial involves four phases of functioning of the architecture: the planning phase, (possibly) the exploration phase, the action execution phase, and the learning phase.

Algorithm 1 Pseudo-code of the model functioning.

```
1: loop VisuoMotorTrials ∈ {1, 2, ..., 120}
2:   if (VisuoMotorTrials ≤ 60) then Goal ← AchieveCorrectFeedback
3:   else Goal ← AchieveIncorrectFeedback
4:   EntropyThreshold ← EntropyMax, Planning ← TRUE, Action = NULL
5:   InitialState ← Observe(Environment)
6:   while Planning do                                  ▷ Planning phase
7:     ForwardSampling(InitialState)
8:     Entropy ← ComputeEntropy(AssociativeLayerActivation)
9:     if (Entropy > EntropyThreshold) then
10:      Planning ← FALSE
11:    else
12:      if (SampledOutcome = Goal) then
13:        Action ← SimulatedAction()
14:        Planning ← FALSE
15:      else
16:        UpdateGoalAssociativeConnections()
17:        LowerEntropyThreshold()
18:  if (Action = NULL) then                          ▷ Exploration phase
19:    Action ← ComputeExplorationAction()
20:  PerformActionInEnvironment (Action)                ▷ Action phase
21:  Outcome ← Observe(Environment)
22:  TrainWorldModel(InitialState, Action, Outcome)   ▷ Learning phase
23:  if (Outcome = Goal) then
24:    UpdateGoalAssociativeConnections()
25:  else
26:    TrainExplorationComponent(InitialState, Action, Outcome)
```

At the beginning of each trial the system observes a colour (lines 5). After the colour observation, the model performs a variable number of 'planning cycles' (planning phase, line 6). During a planning cycle, which lasts 15 steps as the actual trial (as in [49]), the input layer is activated with the observed colour for the initial 5 time steps and then is switched off. As a consequence, the associative-layer neurons fire in sequence to simulate a possible colour-action-feedback concatenation (line 7).

During one planning cycle, the arbitration mechanism operates as follows. The sequential neuron sampling causes a certain activation (membrane potential) of the neurons of the associative layer, encoding the probability over the hidden causes: this probability distribution is used to compute the entropy at each step, and this entropy is then averaged over the sampling steps forming the whole planning cycle. This average entropy is considered as the measure of the uncertainty of the world model (line 8). If this uncertainty is higher than a certain threshold, the arbitration component stops planning as not enough confident on the knowledge of the world model (lines 9-10). Instead, if the uncertainty is lower than the threshold the arbitration component checks if the sampled sequence has produced a state ('read out' in the output layer) that matches the goal (lines 11-14), and if this is the case it stops planning and performs the action in the environment. Instead, if the arbitration component is confident on the world model but the sampling has not produced a sequence that matches the goal, it performs two operations before starting a new planning cycle: first, it updates the goal-associative connections so as to lower the goal-conditioned probability of the wrong sampled sequence (line 16); second, it lowers the entropy threshold of a certain amount to ensure that across the planning cycles the probability of terminating the planning process progressively increases (line 17): this avoids that the model gets stuck in planning. As soon as the planning process terminates, if the model has not found an action that leads to the goal then the action is selected by the exploration component (lines 18-19).

After this, the agent engages again with the environment. In particular, the action selected either by the planning process or by the exploration component is performed in the environment (line 20). Consequently, the environment produces an outcome (correct/incorrect feedback) perceived by the agent (line 21). Based on the observation of the initial state (colour), performed action (finger press), and outcome (correct/incorrect feedback) from the environment, the world model learns (line 22). In particular, it learns the internal representation (hidden causes) of the observations (input-associative connections), the possible time dependencies between them (internal connections of the associative layer), and the generation of the observations (associative-output connections). Moreover, if the performed action has led to actually accomplish the goal in the environment, the goal-conditioned probability of the sampled successful sequence is increased (goal-associative connections; line 24). Instead, if the action failed then only the exploration component is trained to lower the probability of selecting the same action in correspondence to the experienced initial state and goal (line 26).

Note that when a trial starts, the architecture performs a planning cycle to evaluate entropy: this hypothesis is based on the fact that the task is novel. In a more general case where the agent might also encounter familiar tasks a common habit/planning arbitration process might evaluate if a habit is available to solve the task before triggering planning and the planning/exploration arbitration process considered here.

Note also that in case of goal-failure the goal-associative connections are updated during planning to exclude the multiple sampling of the same wrong sequence and action; instead, in the case of goal-achievement such connections are updated after the action is successfully performed in the environment, rather than in simulation during planning: this avoids a training based on the possible false-positive errors of planning (false-negative errors are less likely during planning as the world model learns on the basis of the ground-truth information from the

world). The exploration component is instead trained after the failure of the action executed in the world to avoid to repeat the selection of the actions found to be wrong (this mechanism is analogous to the 'inhibition-of-return' found in visual exploration, leading to exclude from exploration already explored items [54]); the component is instead not trained in case of success as this would amount to habitual learning not possible in few trials. These hypotheses were isolated through the search of the conditions for the correct reproduction of the target human data of the visuomotor task while fulfilling the challenging constraint that planning has to take place while learning the neural world model.

Based on these mechanisms, at the beginning of the visuomotor test the model tends to sample random neuron sequences within the associative layer as the world model has no knowledge on the environment. The arbitration component thus soon passes the control to the exploration component that decides which action to execute, and this is performed in the environment. With the accumulation of experienced trials, the world model improves by learning the hidden causes of observations (colours, actions, feedback) and the time dependencies between them. This leads the arbitration component to measure a higher confidence in the world model, so planning continues and samples, with a higher probability, the (hidden causes of) colour-action-feedback sequences that actually exist in the world. When a planning cycle simulates an action that predicts a goal achievement in the output layer, and the action is actually successful when performed in the environment, this leads to increase the goal-conditioned probability of sampling such sequence again so that the next time the same colour is encountered the sequence is readily selected by the planning process.

1.3 Goal-directed behaviour model: Detailed functioning

1.3.1 The hidden Markov model represented by the world model. This section first illustrates the graphical models corresponding to the *Hidden Markov Models* (HMMs) and the *Partially Observable Markov Decision Processes* (POMDPs) on which planning as inference is grounded, and then explains the particular HMM instantiated by the world model of our architecture. Next the section illustrates the spiking neural network used to implement this world model and links it to the probabilistic aspects of HMMs.

Fig 3 shows a HMM [49, 55] represented through a *graphical model*. A HMM assumes that the agent cannot directly access the world states (they are 'hidden' to it) but only infer them on the basis of noisy observations. In particular, the model represents the world states with a different probability distribution, over the possible *hidden causes*, for each time step. The state probability distribution at each time step is assumed to depend only on the state of the previous time step (*Markov property*); the probability distribution over observations is assumed to depend only on the current state. An agent can use a HMM representing the world dynamics to internally simulate possible sequences of states that the environment might traverse, e.g. to represent the places seen while moving through a corridor or the positions of a displaced object.

Building on HHMs, POMDPs again assume that the agent can access the states of the world only indirectly through noisy sensors (they are 'partially observable') but they also consider the agent's behaviour, in particular the probability distributions of actions at different times. Action probability distributions are conditioned on the internal representations of states (thus forming probabilistic *policies*), and over perceived *rewards*. Rewards are considered as additional observations and assumed to depend on other events such as the world states (different models can make different assumptions on rewards). POMDPs can be used to implement planning by conditioning probability distributions on high rewards (or on a reached goal

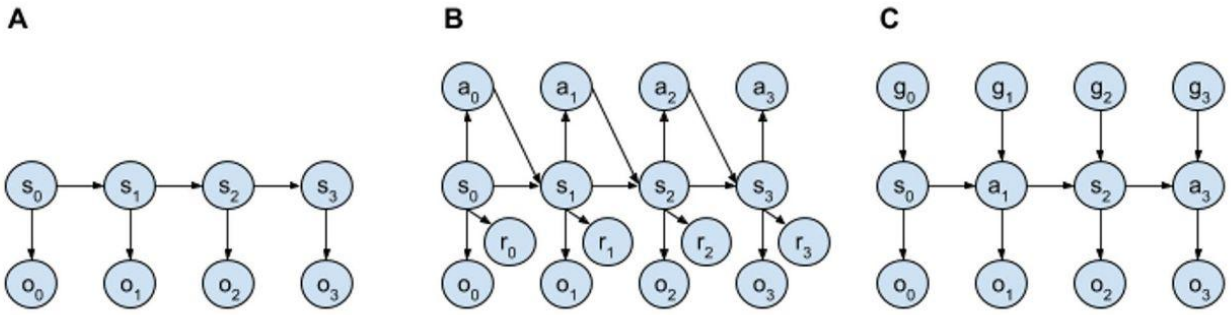


Fig 3. Graphical models of some probabilistic models usable to represent the dynamics of the world in planning systems. Nodes represent probability distributions and directional links represent conditional dependence between probability distributions. (a) Hidden Markov Models (HMMs): these are formed by state nodes 's' and observation nodes 'o'. (b) Partially Observable Markov Decision Processes (POMDPs): these are also formed by action nodes 'a' and reward nodes 'r' (different versions of these models are possible based on the chosen nodes and their dependencies). (c) The HMMs considered here, where the planner knows the currently pursued goal 'g' and observes not only states but also actions (note that the task considered here involves a sequence of independent state-action-state experiences).

https://doi.org/10.1371/journal.pcbi.1007579.g003

state), and then by inferring the probability distributions of the state-action sequences causing them with a high likelihood (*planning as inference*, [24–26]).

A HMM considers the hidden causes of world states, h_t, and observations of them, o_t, as random variables at the time steps $t \in \{0, 1, \ldots, T\}$ forming the sequences $H = \{h_0, h_1, \ldots, h_T\}$ and $O = \{o_0, o_1, \ldots, o_T\}$. The joint probability of these sequences can be expressed and factorised as follows given the assumptions on the probability independencies of the model shown in Fig 3A:

$$p(H, O) = p(h_0) \cdot p(o_0) \cdot \prod_{t=1}^{T} [p(o_t|h_t) \cdot p(h_t|h_{t-1})] \quad (1)$$

This formula highlights the two key elements of the HMM, namely the *generative model* of how the world states (hidden causes) cause the observations, $p(o_t|h_t)$, and the *prediction model* of how a world state causes the following state $p(h_t|h_{t-1})$ (in the neural implementation of the HMM we will equivalently consider the probabilities $p(o_t|h_{t-1})$, and also $p(h_t|o_{t-1})$, to follow the general rule of physical causality for which the current state of any part of the neural network and of the world can depend only on the past state of other parts of the network or the world).

The HMM has parameters θ that are adjusted on the basis of collected data (observations) so that the probability distribution $p(O|\theta)$ converges towards the empirical distribution from the world, $p^*(O)$:

$$\theta^* = \arg\min_{\theta} DL(p^*(O) \| p(O|\theta)) \quad (2)$$

where $DL(.\|.)$ is the Kullback-Leibler divergence between the two distributions and θ^* are the searched optimal parameter values of the model. This problem cannot be solved in closed form and so θ^* are commonly searched numerically, in particular through an *expectation-maximisation* (EM) algorithm. Here we refer to how this is done in versions of HMMs [49, 56] most similar to the neural implementation of HMMs considered here. For these models, the EM algorithm converges towards the solution by alternating an estimation step (E-step) and a maximisation step (M-step): broadly, the E-step samples a sequence of specific values of the hidden causes, H', based on the posterior distribution $p(H|O', \theta)$ dependent on the actual observations O'; the M-step adjusts θ to increase $p(H'|O', \theta)$. In the E-step, the sampling of H'

given O' can be approximated by *forward sampling* [57], i.e. by sampling the h_t distributions in sequence, staring from h_0, given the $\{o'_0, o'_1, , \ldots o'_t\}$ values observed until t.

1.3.2 The spiking neural-network world model. The neural implementation of the world model instantiating the HMM is based on two learning processes. The first process, involving the input-associative connections, learns the hidden causes of different observations as probability distributions of the spikes of the neurons of the associative layer. The second process, involving the connections internal to the associative layer, learns the temporal dependencies between the hidden causes of observations as conditional probability distributions of the spikes of the neurons of the associative layer taking place at succeeding time steps.

The membrane potential of each neuron of the associative layer reflects the activation that would result from the typical connectivity pattern of cortex and other areas of the brain, formed by neurons that reciprocally inhibit each other through inhibitory interneurons. This connectivity pattern tends to keep a constant overall firing rate of the layer. In detail, the membrane potential u_k of a neuron k of the model is:

$$u_k(t) = \hat{u}_k(t) - i(t) \tag{3}$$

where $i(t)$ is the common inhibition received by all neurons caused by the inhibitory interneurons to which they project (this inhibition process is abstracted with a *soft-max* function, see below), and $\hat{u}_k(t)$ is the total activation received from other neurons:

$$\hat{u}_k(t) = \sum_{i=1}^{I} w_{ki} \cdot s_i(t-1) + \sum_{g=1}^{G} w_{kg} \cdot s_g(t-1) + \sum_{a=1}^{A} w_{ka} \cdot s_a(t-1) \tag{4}$$

where w_{ki} are the input-associative connection weights, w_{kg} are the goal-associative connection weights, w_{ka} are the internal associative connection weights, $s_i(t)$, $s_g(t)$, and $s_a(t)$ are the incoming spike signals ($s \in \{0, 1\}$) from the neurons of respectively the input, goal, and associative layer. In the simulations reported in the paper, we also added a Gaussian noise (standard deviation ν) to the membrane potential $\hat{u}_k(t)$ of associative and output neurons to check the robustness of the model: this did not alter the results with respect to the model not encompassing such noise.

We then assume, as in [49, 58], that the firing rate $v_k(t)$ of a neuron k, reflecting its spiking probability, is exponentially dependent on the membrane potential:

$$v_k(t) = \nu \cdot e^{u_k(t)} \tag{5}$$

where ν is a constant scaling the firing rate. This implies the following dependency of the neuron firing rate on the activation from other neurons and on the inhibition from the common inhibition:

$$v_k(t) = \nu \cdot e^{(\hat{u}_k(t) - i(t))} = \nu \cdot \frac{e^{\hat{u}_k(t)}}{e^{i(t)}} = \nu \cdot \frac{e^{\hat{u}_k(t)}}{\sum_{l=1}^{L} e^{\hat{u}_l(t)}} \tag{6}$$

where ν now shows to be the total constant firing rate of the population and $i(t)$ is assumed to be:

$$i(t) = ln \sum_{l=1}^{L} e^{\hat{u}_l(t)} \tag{7}$$

The spiking models we are considering [38, 49] were implemented by assuming continuous time and an inhomogeneous Poisson process to generate the timing of the spikes. However, here we used the version of the model proposed in [38] that considers discrete time steps, a

time-binned binary Excitatory Postsynaptic Potentials (EPSP), and a winner-take-all competition to generate a spike at each step. Although having less biological features, this simpler version of the model facilitates the analyses and derivations and at the same time preserves (and possibly strengthens) the probabilistic interpretability of the spiking networks considered. By assuming $\nu = 1$, Eq 6 becomes a *soft-max* function that abstracts a lateral inhibition-based winner-take-all neural competition. Indeed, now the layer constant total firing is $\sum_{k=1}^{K} \nu(k) = 1$ and $\nu(t)$ can be interpreted as $\nu(t) = p_t(k)$, with $p_t(k)$ being a categorical probability distribution indicating the likelihood that the neuron with index k fires a spike at time t while the other neurons remain silent. Following [49], the neurons also had a refractory period r obtained by subtracting from $u_k(t)$ a value decaying exponentially at each step t ($t = 0, 1, 2, \ldots$) as $r = r_0 \cdot exp\left(\frac{-t}{\tau}\right)$ (where $r_0 = 1.1$, $\tau = 9.5$). This feature revealed very important to allow the emergence of groups of neurons encoding the input patterns as latent causes. The output layer, receiving afferent connections from the associative layer, is formed by a set of neurons behaving as the associative layer neurons.

The weights of the connections linking the input-associative layers, the associative-output layers, and the associative neurons between them, are updated through a Spike-Timing Dependent Plasticity (STDP) rule [59–62]. In particular, we used the following STDP learning rule from [38, 49] to update a connection weight $w_{post,pre}$ linking the pre-synaptic neuron *pre* to the post-synaptic neuron *post*:

$$\Delta w_{post,pre}(t) = \zeta \cdot s_{post}(t) \cdot (e^{-w_{post,pre}} \cdot s_{pre}(t-1) - c) \tag{8}$$

where ζ is a learning rate parameter, $\Delta w_{post,pre}$ is the size of the connection weight update, $s_{post}(t)$ and $s_{pre}(t-1)$ are respectively the spike activations ($s \in \{0, 1\}$) of respectively the post-synaptic neuron in the current time step and the pre-synaptic neuron in the last time step, and c is a constant ($c \in [0, 1]$). The learning rule operates as follows. The rule updates the weight only when the post-synaptic neuron fires ($s_{post}(t) = 1$). When this happens, but the pre-synaptic neuron does not fire ($s_{pre}(t-1) = 0$), then $w_{post,pre}$ decreases of $-\zeta \cdot c$. This leads the post-synaptic neuron to form negative connections with all the pre-synaptic neurons that tend to not fire before it fires. Instead, if the pre-synaptic neuron fires before the post-synaptic neuron ($s_{pre}(t-1) = 1$), then $w_{post,pre}$ increases if $c < e^{-w_{post,pre}}$ and decreases otherwise. This implies (as it can be seen by solving for $\Delta w_{post,pre}(t) = 0$ and setting $s_{post}(t) = 1$ and $s_{pre}(t-1) = 1$) that $w_{post,pre}$ will tend to converge to the positive point $w^*_{post,pre} = -ln(c)$ reached when $e^{-w_{post,pre}} = c$. Overall, for a given neuron the rule thus tends to form positive incoming connections from neurons that fire just before it fires, and negative connections from all other neurons.

The connections that the model learns are the means through which the system implements conditional probabilities. For example, initially the associative units k, each representing possible hidden causes of observations, tend to fire with a certain prior probability distribution, say $p(k)$. The formation of input-associative connections allows an observation i to generate the posterior conditional probability distribution $p(k|i)$ that for example implies an increased probability of selection of the hidden cause k.

Within the associative layer, the learning rule leads to form a connectivity that supports a sequential activation of the neurons encoding the hidden causes of the observations, where the sequence reflects the temporal order in which the observations, reflecting the world states, are experienced by the HMM. The reason is that once the hidden causes are formed, based on the input-associative connections, then they tend to fire in sequence under the drive of the observations. As a consequence, the learning rule leads each associative neuron to connect with the associative neurons that fired before it and to form negative connections with those that did not fire before it. In this way, the connections within the associative network tend to form

chain-like neural assemblies. These connections are hence able to represent the temporal dependencies between hidden causes, for example between a and k corresponding to two succeeding observations, as conditional probabilities $p(k|a)$. Importantly, if the system observes different events following the initial observation of the trial (e.g., different actions and different outcomes after a certain initial colour), the world model will exploit its stochastic neural processes to represent such possible alternative sequences of events. This is at the core of the architecture's capacity to internally simulate alternative courses of actions and events and hence to plan in a goal-directed manner.

The same learning rule is also used to train the associative-output connections. Initially, the output layer expresses a probability distribution, say $p(o)$, that tends to be uniform and so when sampled it generates unstructured observations. With learning, the world model strengthens some connections between the spiking sequences sampled within the associative network and the observations activating the output layer. When the associative-layer world model samples an internal sequence of spikes, this leads to generate the observations on the basis of the reconstruction probability $p(o|k)$.

When the planning process has to generate an action to perform, or a predicted feedback to compare with the goal, the generated event at the output layer is considered to be the one that fired the most during the planning cycle. In cases where the system should generate sequences of events involving multiple actions and predicted states, one should consider other 'reading out' mechanisms, for example one where an event is generated each time the units encoding it fire a minimum number of spikes in sequence.

The goal-associative connection weights are updated on the basis of the failure to achieve the goal during planning and in the case of success when the action is performed in the environment. The weight update is done on the basis of the following reinforcement learning rule:

$$\Delta w_{kg} = \eta \cdot m \cdot ET_k \cdot \left(\frac{w_{max} - |w_{kg}|}{w_{max}} \right) \cdot s_g \tag{9}$$

where η represents the learning rate, m is the pseudo-reward, equal to 1 if the sequence resulted in a successful goal matching (when executed in the environment) and −1 if it resulted in a failure (during planning), ET_k is the Eligibility Trace of the associative unit k, equal to 1 for units that have fired at least once during the planning cycle/trial and to 0 otherwise, and w_{max} is the maximum absolute value that the weight can reach ($w_{max} = 0.5$), and s_g is the activity of a goal neuron. The goal-associative connections allow the goal g to condition the probability distribution over the hidden causes, $p(k|i, a, g)$. With learning, this allows the goal to condition the probability of the sampled hidden causes sequences so as to increase the likelihood of those that involve the correct action. Moreover, when the goal changes, the model is able to modify the conditioned probability of the sequences so as to increase the probability of sampling a different sequence, based on the same world model, achieving the new desired goal.

1.3.3 Arbitration and exploration components. The arbitration component decides if continuing to plan or to pass the control to the exploration component and/or to perform the action selected by either the planning or the exploration process. The component makes these decisions on the basis of a key information, namely an estimation of the level of knowledge of the world model for the given trial depending on the observed colour. This knowledge is related to the fact that the world model has possibly learnt that some sequences of events (action-feedback) might follow the initial observation. A good level of knowledge means that the probability mass of the distribution $p_t(k|i, a, g)$ during the planning cycle steps t is concentrated on few possible hidden causes. The measure of this knowledge at a given time step t can

thus be based on the entropy of the probability distribution expressed by the associative layer:

$$H_t(k|i,a,g) = -\sum_{k=1}^{K} p_t(k|i,a,g) \cdot ln(p_t(k|i,a,g)) \quad (10)$$

where the maximum value of such entropy is $H_{max} = ln(K)$ corresponding to a uniform probability distribution where each k neuron of the layer has the same probability of firing $p(k) = 1/K$. The measure of the uncertainty H of the world model in a given planning cycle lasting T time steps is in particular defined as:

$$H = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{H_t(k|i,a,g)}{H_{max}} \right) \quad (11)$$

At the end of each planning cycle, the arbitration component computes H, compares it with an entropy threshold $H_{Th}(t)$, compares the action-outcome z with the pursued g, and selects one of three possible functioning modes of the architecture:

- $H < H_{Th}(t)$ and $z \neq g$. The goal-associative connections are updated and a new planning cycle starts.
- $H < H_{Th}(t)$ and $z = g$. Planning stops and the action of the last planning cycle that caused the anticipation of the goal is executed in the world (without activating the exploration component).
- $H_{Th}(t) < H$. Planning stops and the exploration component selects the action to perform.

The entropy threshold decreases linearly with each planning cycle so that the exploration component is eventually called to select the action if the planning process fails to reach the goal multiple times:

$$H_{Th}(t) = \epsilon - (f \cdot \delta) \quad (12)$$

where ϵ is the value to which the entropy threshold is set at the beginning of the trial (and the planning process), δ is its linear decrease, and f is the number of failed planning cycles performed in the trial.

The exploration component is a neural network formed by two layers. The first is an input layer formed by 6 neurons encoding the elements of the Cartesian product between the possible 3 colours and 2 goals. The second is an output layer formed by 5 neurons representing the possible actions, receiving all-to-all connections from the input layer. When the exploration component is called to select the action, the input layer is activated according to the current colour-goal combination (hot-vector activation), the activation potential of the second layer units is computed as usual as the sum of the weighed inputs, and an action is chosen on the basis of a *soft-max* function (Eq 6). When the action leads to a negative reward (−1, received in case of goal missed), the connection weights of the component are updated using the same reinforcement learning rule used for the goal layer (Eq 9). This tends to exclude actions that are not useful for the current state-goal combination, thus fostering exploration. Note that an additional slow-learning component similar to the exploration component might be used to model the formation of habits in experiments involving long learning periods.

1.4 Search of the model parameters

The model functioning depends on seven important parameters, indicated in Table 1. We searched the best values of those parameters by fitting the model behaviour to the

Table 1. Parameters identified with the grid search technique. In particular, parameter names, minimum and maximum range, and values found by the search.

Name	Range min	Range max	Found value
STDP learning rate (ζ)	0.1	1.0	0.96
STDP threshold (c)	0.1	1.0	0.67
Planner learning rate (η)	0.001	1.0	0.008
Softmax temperature (τ)	0.01	0.1	0.02
Neural noise (ν)	0.01	0.1	0.02
Entropy max threshold (ϵ)	0.3	1.0	0.74
Entropy decrease (δ)	0.01	0.2	0.12

https://doi.org/10.1371/journal.pcbi.1007579.t001

corresponding data of the human participants. In particular, we randomly sampled and evaluated 100,000 parameter combinations. For each combination, we recorded and averaged the behaviour of 20 'simulated participants', in particular their performance in the 20 trials for each of the stimuli S1, S2, and S3, and the average over colour of the reaction times (this because the original data on the reaction times of humans were not separated). Such three performance datasets and one reaction-time dataset were compared with the corresponding average data from 14 human participants through a Pearson correlation coefficient $R_{d,m}$ computed as:

$$R_{d,m} = \frac{C_{d,m}}{\sqrt{V_d * V_m}} \tag{13}$$

where $C_{d,m}$ is the covariance between the data from humans, d, and data from the model, m, and V_d and V_m are their respective variances. In particular, the coefficient was computed separately for the different data sets (performances and reaction times) and then averaged.

The range of the parameters explored by the search, and the best parameter values that it found, are shown in Table 1. The best parameter values, that had a correlation coefficient of 0.72, were used in all the simulations illustrated here.

2 Results

This section illustrates the behaviour and functioning of the model when tested with the visuomotor learning task proposed in [15] and described in Sec 1.1.

2.1 Behavioural analysis

Fig 4 shows that the model exhibits a performance similar to the human participants by comparing the probability of correct responses in repeated trial triplets for 14 human participants (from [15]) and 20 simulated participants (obtained with different seeds of the random-number generator). The performance of the model is similar to the human one for stimuli S1 and S2 whereas it is slightly lower for S3. Once the model finds the correct action for a stimulus, when it encounters such stimulus again it reproduces the correct action with a high probability similarly to the humans. The architecture takes more cycles to converge to such a high probability for S3 than for S1 and S2 because the planner has a larger number of wrong sequences to explore and so has a higher probability of wrongly anticipating a correct feedback. This problem is less impairing for S1, and in part for S2, involving fewer wrong sequences during planning.

Fig 5 compares the reaction times of the same human participants (from [15]) and the simulated participants considered in Fig 4. The reaction times of the model are produced by these

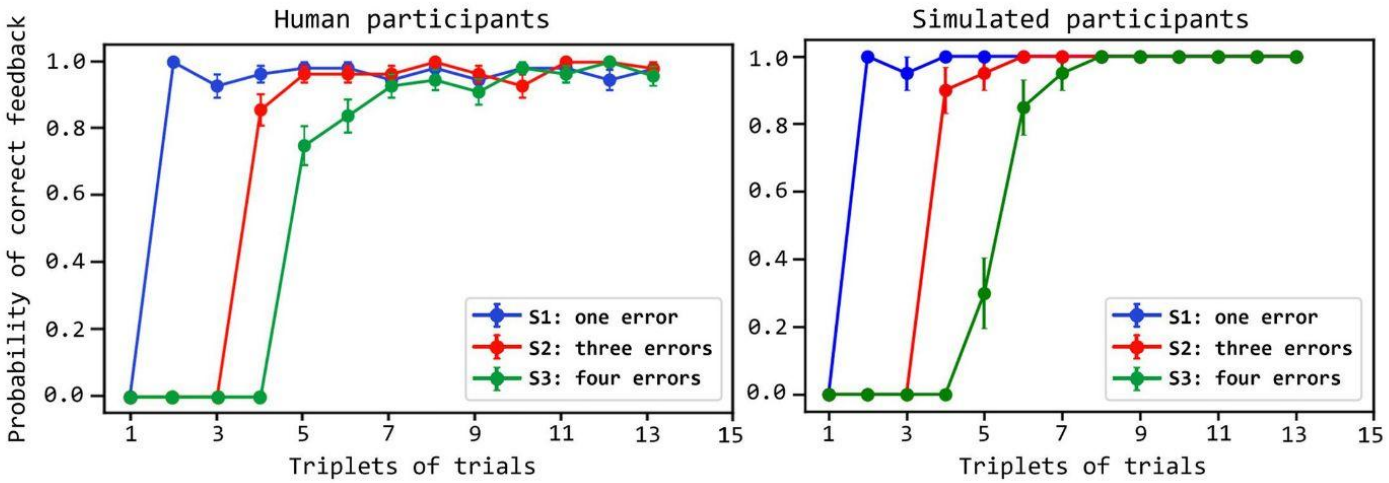


Fig 4. Comparison of the performance of the human and simulated participants. The performance (y-axis) is measured as the proportion of correct feedback over the trial triplets (x-axis), plotted separately for the three different colour stimuli (S1, S2, S3). Curves indicate the values averaged over 14 human participants and 20 simulated participants; error bars indicate the standard error. The data of human participants are from [48].

https://doi.org/10.1371/journal.pcbi.1007579.g004

processes. The arbitration component decides to implement a different number of planning cycles, each involving the generation of colour-action-feedback sequences, depending on the knowledge stored in the world model. If a larger number of planning cycles is performed, this results in longer reaction times. As shown in the graph, the reproduction of the human reaction times is particularly interesting and challenging as it has an inverted 'U' shape. This shape is reproduced and accounted for by the model on the basis of the following processes. In the first trials, for each stimulus the entropy (uncertainty) of the world model is high as the associative layer expresses a rather uniform probability distribution. Indeed, the component has

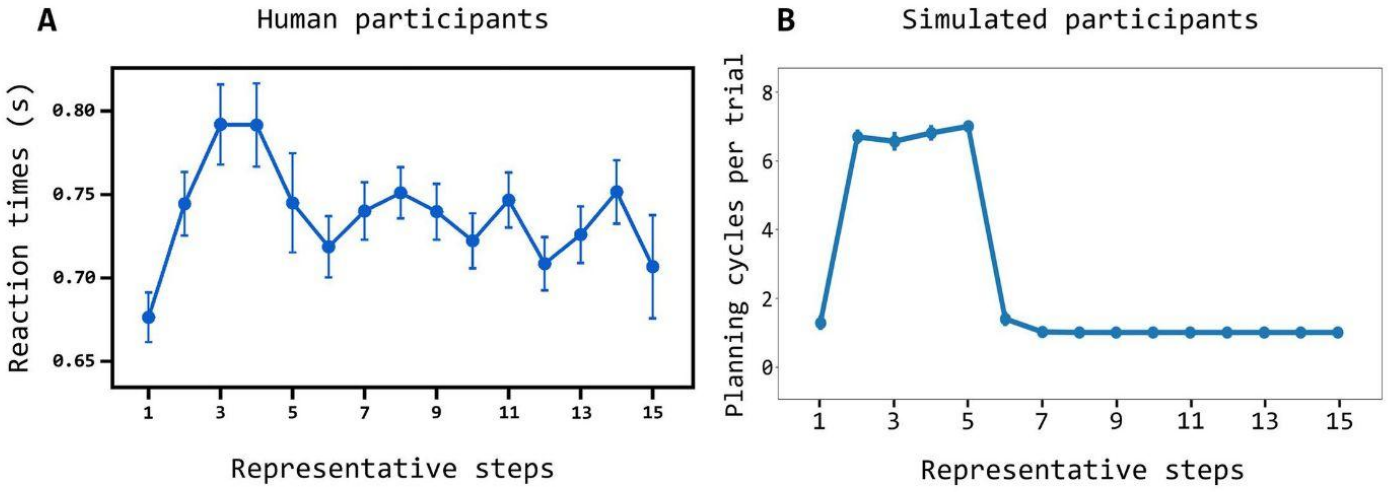


Fig 5. Comparison of the reaction times of the humans and simulated participants. (A) Reaction times of human participants averaged over S1, S2, and S3 (y-axis) for the 'representative steps' ([48]; x-axis); the 'representative steps' allow the alignment of the reaction times of the three stimuli so as to separate the exploration phase (first 5 steps) and the exploitation phase (6 steps onward); to this purpose, the reaction times for S1 obtained in succeeding trials from the first onward is assigned the steps (used to compute the averages shown in the plot) '1, 2, 6, 7, . . .', whereas S2 is assigned the steps '1, 2, 3, 4, 6, 7, . . .', and S3 is assigned the steps '1, 2, 3, 4, 5, 6, 7, . . .'; data are taken from [48]; (B) Reaction times of the model, measured as number of planning cycles performed in each trial, plotted in the same way as done for humans. Error bars indicate mean standard errors.

https://doi.org/10.1371/journal.pcbi.1007579.g005

still to identify the hidden causes of stimuli and actions, so the neurons encoding them tend to spike at a similar rate. As the entropy is high, the arbitration component tends to soon pass the control to the exploration component and so the reaction times are low in the initial trials. In the following trials the world model forms representations of the experienced colour-action-feedback sequences and so it assigns to them a higher posterior probability with respect to other patterns. The arbitration component thus tends to compute a lower entropy, the architecture plans for longer, and the reaction times get longer. During this planning process, the associative component tends to sample the learnt sequences with a high probability conditioned to the observed colour. If none of the sequences leads to predict an event that matches the pursued goal through the output layer, the probability of such sequences is however decreased under the conditioning of the goal; the control is thus passed to the exploration component. When, during these trials, the action performed in the world manages to produce the desired goal, the world model learns the corresponding sequence and assigns to it a high posterior probability. When the colour of such sequence is observed again, the sequence is sampled with a higher probability and results in a successful outcome-goal match. The arbitration component stops planning and the action is performed in the world. The reaction times thus become short again. Overall these processes reproduce the inverted 'U' shape of the reaction times similar to the one observed in humans.

2.2 Model internal dynamics

Fig 6 shows how the activation of the associative layer during planning triggered by the different colours evolves across the succeeding trials of the test due to the increasing knowledge acquired by the world model and by the goal bias. In the initial phases of learning (trials T1-T3 for S1, S2, and S3), the prior probability of activation of the neurons of the associative layer tends to be quite uniform, thus resulting in a random spike sampling of the neurons still not encoding in a sharp way the hidden causes of different colours, actions, and outcomes. This means that the model has not yet identified specific hidden causes of the observations, the temporal relations between them, and the correct colour-action-feedback sequences associated to the three colours.

Based on the observations of the world, the STDP rule acting on the input-associative connections leads the associative layer (world model) to form an internal representation of the hidden causes of the observations, namely of the colours, actions, and feedback. At the same time, the STDP plasticity internal to the associative layer leads it to form a HMM that represents in an increasingly accurate fashion the time-related probabilistic dependencies between the discovered hidden causes. Finally, while possible sequences are encoded by the associative layer, starting from the observed colour, the STDP acting on the goal-associative connections progressively increases the probability of sampling sequences that lead to achieve the goal and to decrease the probability of the sequences that fail to do so.

The effect of these plasticity processes can be seen in the figure graphs (Fig 6). With respect to S1 (three graphs at the left), a population of neurons encoding the correct colour-action-feedback emerges during the initial trials (T1-T3 graph) and later manifests with a sharp activation (T4-T15 graph). For colours S2 and S3 (respectively second and third column of graphs) a successful population of neurons encoding the correct colour-action-feedback takes longer to emerge: during trials T4-T15 (see related graphs) various neural populations fire with a certain probability, and only in trials T16-T20 one stable population encoding the correct sequence linked to the colour emerges. Importantly, during these learning process the world model, which tends to record any aspects of the world dynamics independently of the fact that it is useful to pursue the current goal or not, also learns sequences leading to an

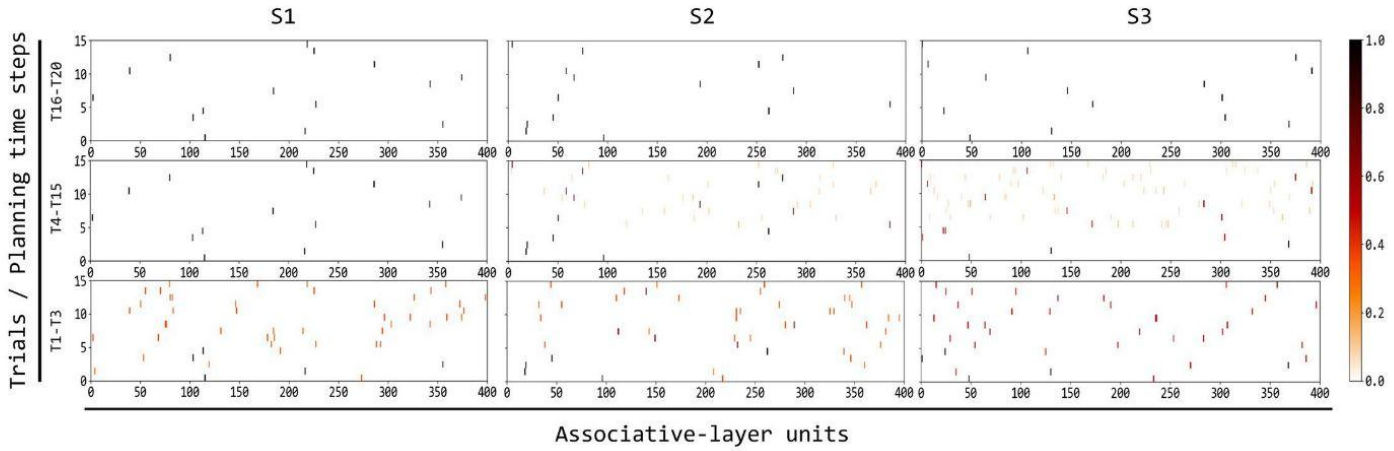


Fig 6. Evolution of the spiking activity of the associative layer units while planning, across the experiment trials. To best interpret the figure, recall that: 15 planning cycles formed one planning sequence (forward sampling), a variable number of planning sequences was generated in one trial for a given colour, 3 trials for the different colours formed a triplet, 20 triplets formed the whole test. The figure shows data collected while the model planned during the trials of the experiment related to each colour, from trial one (T1) to trial 20 (T20). Each column of graphs corresponds to a different colour stimulus, respectively S1, S2, and S3. For each of the nine graphs, the x-axis indicates the indexes of the 400 neurons of the associative layer, and the y-axis indicates the 15 planning cycles of the planning sequences produced in each trial (in each graph the planning cycles progress from bottom to top). Each graph in particular reports the spikes of each neuron for multiple trials (T1-T3 for the bottom row of graphs, T4-T15 for the middle row, and T16-T20 for the third row) and for the multiple planning cycles of those trials: the colour of each little line indicates the proportion of spikes of the corresponding neuron during those trials and cycles.

https://doi.org/10.1371/journal.pcbi.1007579.g006

incorrect feedback. The next section shows how this knowledge might become useful to accomplish other goals.

Fig 7 shows, with analogous graphs, how the activation of the output layer during the planning trials evolves in time due to the increasing knowledge acquired by the world model. The

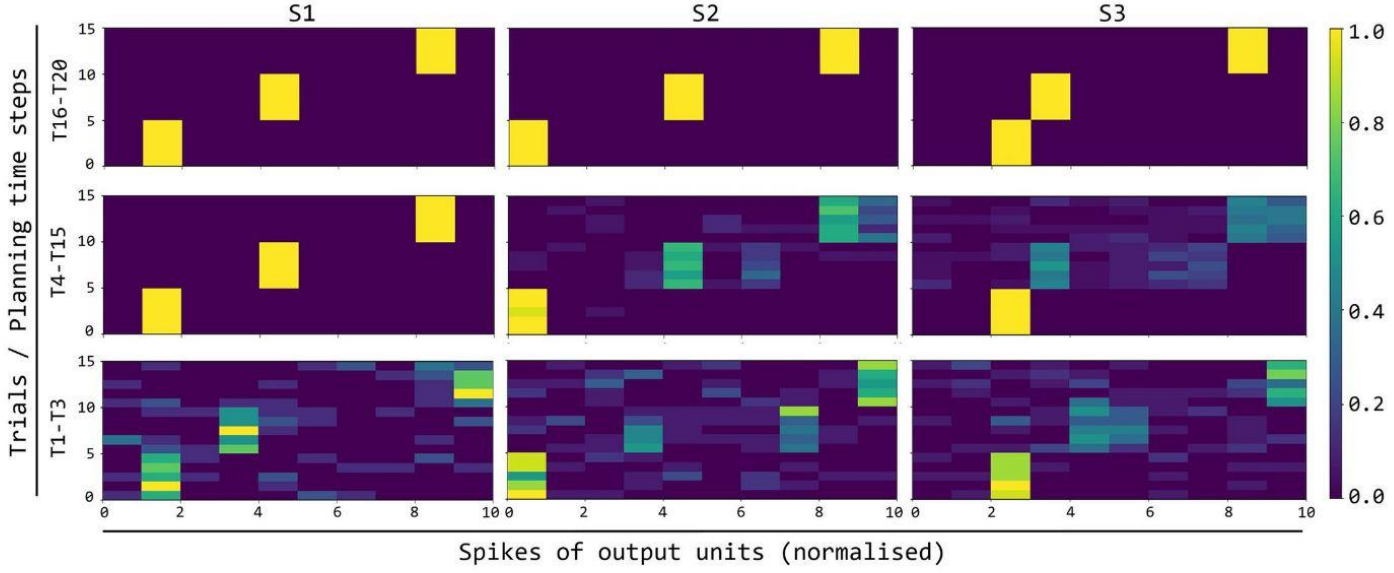


Fig 7. Evolution during trials of the activation of the output layer units encoding the predicted observations and actions. The three columns of graphs refer to the three colour stimuli; the three rows of graphs correspond to different succeeding sets of trials of the task (T1-T3, T4-T15, T16-T20). Each of the nine graphs shows the activation of the 10 output units (x-axis: units 1-3 encode the three colours, units 4-8 encode the 5 actions, and units 9-10 encode the correct/incorrect feedback) during the 15 steps of each trial (y-axis). The colour of the cells in each graph indicates the activation (normalised in [0, 1]) of the corresponding unit, averaged over the graph trials (e.g., T1-T3) and the planning cycles performed within such trials.

https://doi.org/10.1371/journal.pcbi.1007579.g007

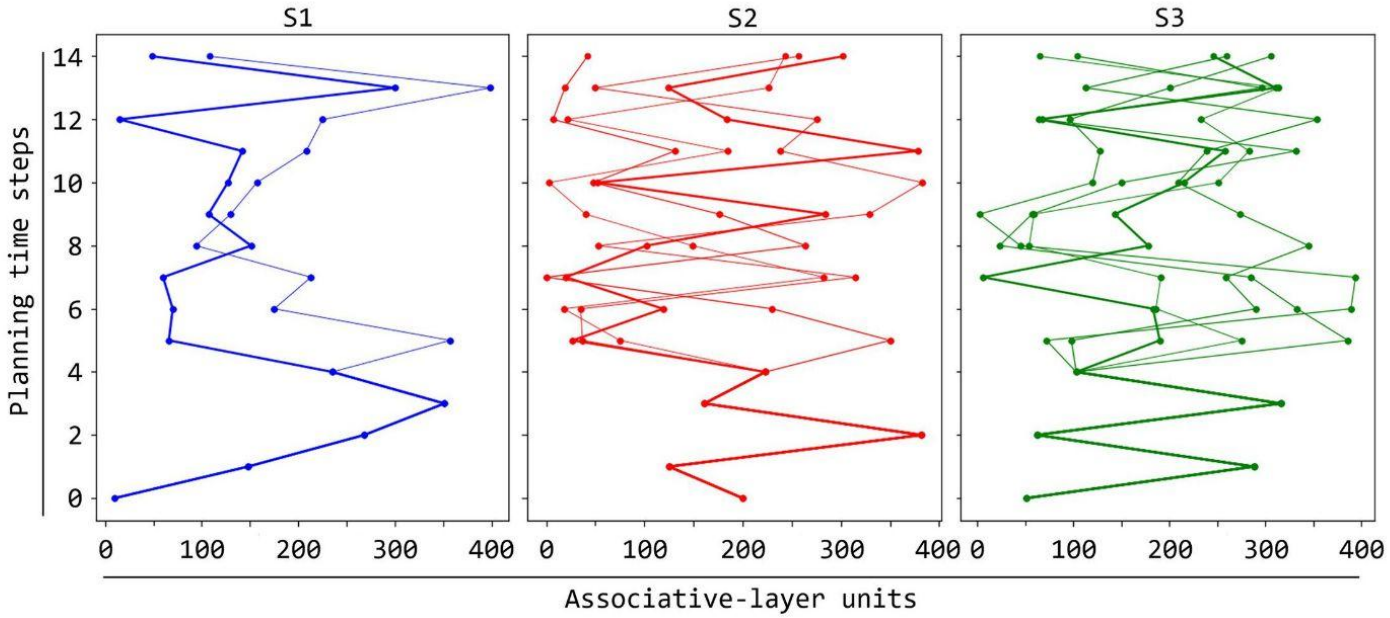


Fig 8. Possible neural trajectories simulated by the model during planning. The three graphs show different neural trajectories that the associative component can generate for respectively the three colours S1, S2, and S3. For each graph, the x-axes indicates the associative neurons and the y-axis the planning time steps and a dot indicates that the corresponding neuron was active. The bolder curve within each graph marks the correct trajectory for the pursued goal 'correct feedback'.

https://doi.org/10.1371/journal.pcbi.1007579.g008

firing of the output layer during planning expresses the predictions of the events (colours, actions, and feedback) that might happen starting from the observed trial colour. Such predictions are based on the simulation of the possible evolution of the world events based on the HMM instantiated by the associative layer. Regarding S1 (left three graphs of the figure), during the first trials (T1-T3) the world model has no or little knowledge on the dynamics of the world, and so the activation of the units in the output layer reflect a uniform probability distribution leading to random predictions of the trial events. With additional experiences of trials involving S1 (T4-T15), the world model starts to learn to represent the trial events and, under the conditioning of the current goal, to assign a high probability to the correct colour-action-feedback sequence. As a consequence, the probability distribution of the output layer starts to correctly predict such correct sequence.

During trials T4-T15 and T16-T20 the same process happens for the correct sequences of the two colours S2 and S3. Also for these stimuli, towards the end of all trials (T16-T20) the probability distribution expressed by the output layer, conditioned to the associative layer activation, has converged to a probability close to 1 for the correct sequences.

Fig 8, showing the neurons of the associative layer spiking in sequence during repeated planning cycles, demonstrates how *emergent generativity* (Sec 'Introduction') allows the model to imagine different possible future action-outcome trajectories in correspondence to the same colour stimulus. The figure also highlights the trajectory that leads to match the 'success' goal. To collect the shown data, we let the model learn until trial T7 for each colour to ensure that it could learn several possible trajectories for it. After this training, we turned off the goal layer to avoid any bias of the associative layer, and let the model perform 400 planning trials for each colour. In this condition, the associative layer responds to the *same* colour by triggering the spikes of *different* possible neuron sequences encoding different colour-action-feedback sequences. Importantly, the figure shows how, when the simulation reaches a 'branching

point' after the activation of the neurons encoding a certain colour, the stochasticity of neurons at the low-level is amplified by the competition between rival neural populations at a higher-level and this allows the model to imagine different possible actions to perform and feedback to receive. This generativity process supports the 'cognitive' exploration of different possible colour-action-feedback trajectories possibly resulting in a successful matching of the 'correct feedback' goal.

2.3 Predictions of the model

An important advantage of planning is that the world model can store general knowledge on the dynamics of the world and this can be used to accomplish different goals. This is a prototypical feature of goal-directed systems that allows them to rapidly switch behavior if the goal changes (*behavioral flexibility*). It was thus interesting to check to which extent the current architecture preserved this capacity since it incrementally acquires a *partial world model* while solving the visuomotor task ('partial' as the solution of the visuomotor task requires the discovery of only the correct colour-action-feedback sequence for each colour, not of all possible sequences). To this purpose, after the architecture solved the task as reported in the previous section, it was given 20 additional trial triplets to pursue the different goal of 'obtaining an incorrect feedback' in correspondence to the three colours. Fig 9A shows the results of this test. When the goal is switched, the architecture is able to rapidly change behaviour and choose the sequences that lead to the desired new goal given the colour. What happens is indeed that, under the conditioning of the observed colour, the world model already has the representations of (a) the hidden causes of the possible observations and (b) of the possible sequences with which such observations might be experienced. In particular, since the previous goal unit is now off, the probability of the different sequences tends to be similar, and so the system tends to sample all of them equally during planning. However, this allows the architecture to rapidly discover a sequence that leads to the desired new goal and thus to increase the probability of generating such sequence conditionally to the new goal.

Regarding the reaction times (Fig 9B), the model shows a transient increase of their size in correspondence to the goal switch. This is due to the fact that with the new goal the architecture needs to perform the sampling of some sequences before finding the successful ones. The reaction time is higher for S1 than for S2 and S3 as for it the model has less sequences available

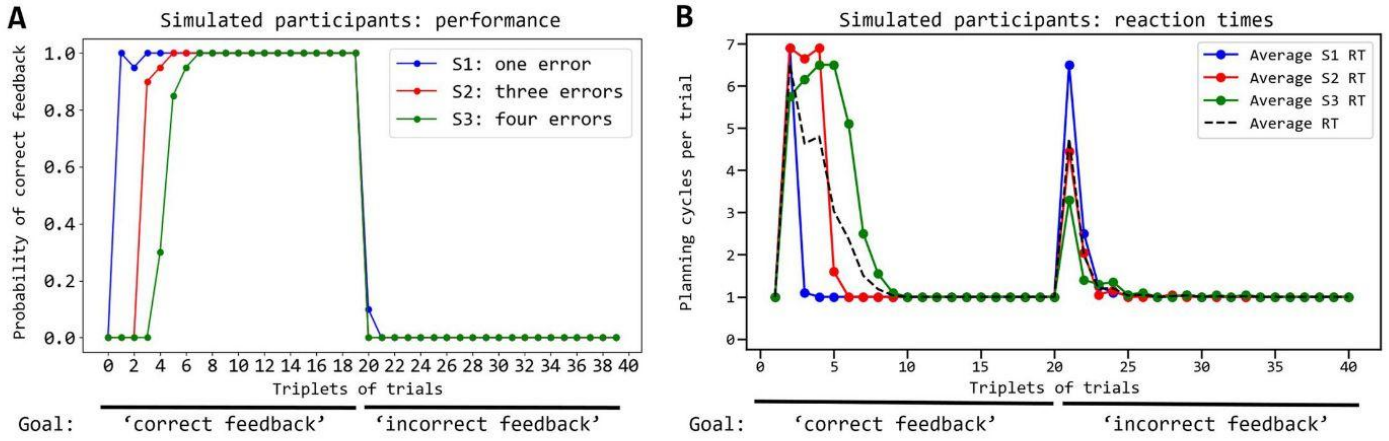


Fig 9. Behaviour of the system when the goal is switched to a new one, averaged over 20 simulated participants. (A) Performance, averaged over the simulated participants, measured as probability of selection of the correct action (y-axis) along the trial triplets (x-axis); the pursued goal is switched from getting a 'correct feedback' to getting an 'incorrect feedback' at triplet 20. (B) Average reaction times measured during the same experiment shown in 'A'.

https://doi.org/10.1371/journal.pcbi.1007579.g009

to reach the new 'incorrect feedback' (instead, the model has exactly one sequence to achieve the 'correct feedback' goal for each colour).

These results represent a prediction of the model that might be tested in a future experiment with human participants through a test analogous to the one presented in this section. In particular, the test should measure the dynamics across trial triplets of the performance and reaction times differentiated by the three colours (S1/S2/S3), as shown in Fig 9.

3 Discussion

This section discusses the main features of the model by first analysing the results presented in the previous section and then by considering the model general features by also comparing it with the current state-of-the-art probability-based or spiking neural-network models of goal-directed behaviour and planning.

3.1 Discussion of the specific results

As shown in Sec 2.1 and in Fig 4, once the model finds the correct action for one colour the probability of correct answers rises steeply, in agreement with what happens with the human participants of the target experiment. Moreover, as in humans, the architecture takes more cycles to converge to such a high probability for S3 because the planner has a larger number of wrong trajectories to explore and so has a higher probability of wrongly anticipating a correct feedback; this problem is less impairing for S1 and S2 involving fewer wrong trajectories that the planning process has to consider. Note how this result, and the one on the reproduction of the reaction times commented below, is not a mere fitting exercise as the architecture reproduces the target data while satisfying a number of biological and ecological constraints, in particular: (a) it solves the task only through goal-directed process, and not through habitual mechanisms as done by previous models of the task [48]: this is requested by the short duration of the task that does not allow habit acquisition; (b) the world model, representing the core of the architecture, relies on spiking-neural mechanisms and biologically plausible circuits; (c) planning takes place while the world model is being acquired, as imposed by the solution of tasks involving new portions of the environment; (d) the model uses an unsupervised learning process.

Fig 5 shows how the model reproduces human reaction times quite accurately as the differences with the target human data are due to some simplifying assumptions of the model. The differences are that the reaction times are above zero for the human participants and close to zero for the model, and that in the first trial they are lower than those of the steady-state trials for the human participants whereas they are similar for the model. The reasons of the first difference is that human participants are likely endowed with an additional habitual/goal-directed arbitration mechanism making a decision before the second exploration/exploitation arbitration mechanism considered here is activated, and this increases the reaction times of a certain amount for all trials. The second difference could be explained by the fact that human participants listen to an explanation of the task before solving it and so they likely start the test having already decided that they should not plan in the first trial, which thus has a low reaction time; instead, the model attempts to plan also in the first trial to check if it is able or not to solve the task.

A second result is the model capacity to reproduce the inverted 'U' shape of the reaction times exhibited by human participants and to explain it. In particular, the model suggests that: (a) in the initial trials, the world model has learnt no state-action sequences, its entropy is high, and so the arbitration component passes the control to the exploration component: the reaction times are hence short; (b) when the world model has learnt some sequences, but these

are wrong, planning implements several cycles to explore such sequences and to lower their goal-conditioned probability, so the arbitration component takes time to pass the control to the exploration component: the reaction times are hence long; (c) when the world model has learnt the correct sequence, entropy is low and thus the planning process samples the correct sequence with a high probability, obtains a successful matching of the goal, and triggers the performance of the related action: the reaction times hence become short again. Another model [48] used an entropy-based measure as a means to decide to give control to a goal-directed component or to a habitual component, and reproduced the 'U' shape of the reaction times observed in the target experiment considered here. This model was based on a goal-directed component formed by a *Bayesian Working Memory* (a memory of the probabilities of the time-dependent states, of the one-step environment transitions, and of the rewards) and a habitual component (based on *Q-learning*). The model reproduced the 'U' shape of reaction times as the sum of two values: (a) the logarithm of the number of items in working-memory, related to the performed trials; (b) the entropy of the action probabilities. The inverted 'U' shape of reaction times was obtained by the fact that 'a' tends to increase with the accumulation of items in memory while 'b' tends to decrease with the diminishing variance of the action probability. In comparison, the model presented here produces the inverted 'U' shape as an emergent effect of the change of knowledge of the world model. The empirical and computational implications of the two hypotheses presented here and in [48] deserve further investigations.

Figs 6, 7 and 8 visualise the internal functioning of the model, in particular the activation of neurons that dynamically encode multiple sequences of colour-action-feedback during planning. The figures highlight two relevant features of the model, in particular: (a) its capacity to autonomously form neural internal representations (hidden causes) of the observations at different times and to activate them in sequence: this capacity relies on the used STDP unsupervised learning rule and the features of the model architecture; (b) the sampling done by spikes of the probability distributions expressed by the world model, and the emergent generativity of the architecture (further discussed below): these processes rely on the stochastic nature of the model, allowing it to 'imagine' different possible action-feedback sequences in correspondence to a colour stimulus.

Fig 9 shows that once the world model has acquired goal-independent knowledge on the environment dynamics, the architecture can use it to pursue different goals 'on the fly', i.e. without the need to further train the world model. This feature is the hallmark of the flexibility of goal-directed behaviour and is shared with the previously state-of-the-art planning-as-inference models relying on spiking neural-networks [39, 42]. However, these models were not used to produce specific empirical predictions as here.

3.2 Discussion of the general features of the model

The results also highlight the novelties of the proposed architecture with respect to the current models. A first novelty with respect to the previous models implementing planning as inference based on brain-like mechanisms [38–40] is that our architecture proposes an hypothesis on how organisms might learn the world model while using it for planning. This is a key challenge for planning, as recently highlighted in [46]. The challenge is different from the exploration/exploitation issue in model-free models [4], and requires arbitration mechanisms different from the classic ones used to balance goal-directed and habitual processes [47, 48]. The work [46] highlights how the challenge is made hard by the 'bad-bootstrapping' problem, mentioned in the Introduction, for which the world model tends to prematurely converge to sub-optimal solutions due to the biased selection of actions directed to pursue goals. The work

also presents a model offering a solution to the challenge based on a suitable balance of the selection of goal-directed and exploration actions. The solution is based on the minimisation of 'free energy', which is the pivotal quantity of the *active inference* framework [22], and in particular on probability functions related to the events of interest and a derivative-based optimisation algorithm. On this respect, our model is the first to present a solution to the challenge based on brain-like spiking neural network. In addition, the model presented in [46] gives a principled solution with respect to previous probabilistic models [47, 48], but for now it is applicable only to simple tasks that require the agent to learn the probability function parameters while being given *a-priori* the set of possible hidden causes of observations. Instead, the model presented here is able to autonomously learn the hidden causes of observations based on spiking neural-network mechanisms.

A second novelty of our model with respect to previous models implementing planning as inference based on brain-like mechanisms [38–40] is that it learns the world model on the basis of a biologically plausible *unsupervised* learning mechanism rather than on the basis of a supervised learning algorithm [38, 39] or by using a world model given *a-priori* [40]. This is an advancement for the biological plausibility of planning as inference models. Indeed, from a computational perspective finding the conditions for the successful functioning of such unsupervised learning process, contextually to the solution of the previous problem related to the acquisition of the world model during planning, represented the hardest challenge found in developing the architecture. We now briefly discuss the three main innovations that support the solution.

First, we grounded learning on the STDP unsupervised learning rule proposed in [49]. This rule is ideal to allow the self-organisation of the architecture associative layer leading to form both the neural representations of hidden causes of observations and the temporal dependencies between them, as required by the autonomous learning of the world model through the spiking recurrent network. Given a neuron that fires, the rule tends to increase the afferent connection weights from neurons that have fired in the recent past, and to decrease connection weights from neurons that have not fired: in the presence of a strong lateral inhibition installing a competition between neurons, as it happens in several parts of brain, this mechanism leads to the emergence of cell groups that specialise to maximally respond to specific (possibly delayed) input patterns. Notice how this mechanism has interesting analogies with the learning processes used in rate-based Self-Organising Maps [63, 64].

A second novel feature that allowed the architecture to autonomously learn the world model is the use of a HMM having a relevant difference with respect to those used in other planning-as-inference spiking network models [38, 39]. These models use a world model based on a classic HMM reproducing possible sequences of states but not actions. Instead, the world model used here is based on a HMM that observes sequences of states *and of actions*, respectively produced by the environment and by another component of the architecture (e.g., by the exploration component used here). This has various possible advantages. One advantage, employed here, is that the world model can directly select actions to perform; instead, previous models [38, 39] need an additional mechanism selecting actions on the basis of the state sequence produced by the world model. A second advantage is that for each environment state the world model can suggest the selection of actions that have a potential relevance in that context, rather than any action (this captures the popular idea of *affordance* in cognitive science [65, 66]). A last advantage could be the easier learning (and understanding) of state-action sequences directed to a goal produced by other agents; indeed, the world model would be neutral with respect to the fact that actions are performed by another part of the brain or by another agent.

A third and last novel feature that allowed the architecture to autonomously learn the world model is the explicit representation of the *goal* used to condition the probability distribution expressed by the world model. Previous state-of-the-art models [38, 39] conflated the goal, initial state, and environment conditions into a whole 'context' representation. Our representation of goals allows their manipulation independently of other conditions, as shown by the model's capacity to successfully plan how to reach new goals on the basis of the experience that the world model acquired in other tasks. Moreover, it paves the way to the enhancement of the architecture with mechanisms allowing the autonomous selection of goals.

We now consider what we think to be a very important mechanism used by the model: emergent generativity. Although shared with other previous models, here we aim to explicitly identify its general features and to stress its wide scope and importance. With *emergent generativity* we refer to the property of a spiking neural-network system for which the low-level stochastic events represented by the spikes of neurons are possibly 'amplified' by the neural circuitry of the system to actively generate multiple alternative high-level patterns –encoding cognitive contents such as percepts, motivations, thoughts, actions, and plans– useful to support adaptive behaviour. The key 'ingredients' of emergent generativity are hence: noisy low-level stochastic units, circuits supporting competitive activation mechanisms, STDP-like unsupervised learning processes, and high-level cognitive processes and behaviours.

Emergent generativity is characterised by two relevant elements. The first element regards 'generativity' and involves the stochastic nature of spike sampling that allows the production of *alternative* patterns in correspondence to the *same* input/context. This process is important as the generation of alternative plausible patterns is at the core of search algorithms possibly employed by brain (by 'plausible patterns' we mean patterns having a high chance to satisfy some constraints, e.g. 'images you might see in a certain environment', or 'actions you might be able to perform with your body'). For example, generativity can support the search of different courses of action that might lead to a desired goal state starting from a given initial condition. In neural networks, generativity is often based on stochastic elements supporting the generation of novel plausible patterns. Notable examples of these systems are Generative Adversarial Networks (GANs; [67]) and Variational Autoencoders (VAEs; [68]) able to generate new plausible input patterns by drawing sample patterns from prior probability distributions of 'latent variables' (hidden causes) and then by transforming them through deterministic neural components trainable with supervised learning (some recent versions of VAEs are also able to learn and generate sequences of hidden causes, analogously to HMMs [69]). These neural systems offer a good intuition on the potential utility of generativity, but within them what we can call the 'stochastic generative engine' (meaning the stochastic mechanism at the core of the generation of alternative plausible patterns) is limited to a particular portion of the system, for example the stochastic input sent to the 'generator' in GANs or the stochastic 'bottleneck' in VAEs. Importantly, such stochastic mechanisms are in contrast with the use of gradient descent algorithms needed to implement supervised learning as they introduce discontinuities preventing differentiation (e.g., VAEs have to use a 'reparameterization trick' to allow the gradient information to 'pass through' the bottleneck stochastic nodes). Instead, in spiking neural networks each spiking neuron, if endowed with intrinsic stochasticity, represents a 'micro stochastic generator' and so the 'stochastic generative engine' of the whole system is distributed in each part of the system rather than being confined in specific locations of the architecture as in GANs and VAEs (as mentioned in the Sec 'Introduction', this shares analogies with *particle filters* [36, 37]). Although this possibility was not exploited here, it might be explored in future work, for example to support planning at multiple levels of abstraction. The use of stochastic units in all parts of the system requires the use of learning rules not requiring differentiation across neural layers, such as the STDP unsupervised

learning rule used here. In this respect, *Boltzmann Machines* and *Restricted Boltzmann Machines* are interesting neural network models that, although now less popular, might be relevant to study systems exhibiting emergent generativity since they are based on an architecture fully based on stochastic neurons and use local unsupervised learning rules [70–72]. The second important element of emergent generativity regards 'emergence' and involves the process for which in complex systems, such as the brain, the dynamical interaction of low-level elements can give rise to organised patterns at higher levels [73]. In particular, in the brain, events involving spike neurons at a low (micro) level are amplified by neural mechanisms in order to generate patterns that encode content, such as perceptions, thoughts and actions, at a higher (macro) cognitive level. As shown here, the 'amplification' can for example rely on circuits implementing winner-take-all competitions grounded on typical connectivity patterns of the brain micro-circuits, and on unsupervised learning processes relying on the brain spike-timing dependent plasticity (STDP) [49, 74–76]. Interestingly, as mentioned above, these mechanisms are analogous to those used in self-organising neural networks [63, 64]. Importantly, the fact that multiple levels of organisation indeed characterise models as the one presented here becomes apparent if one considers that the support of the probability distributions in spiking networks correspond to the identity of neurons, whereas the support of the probability distributions of percepts, actions, and thoughts corresponds to the states of sensors, actuators, and other neural components. This contrasts with the generativity of standard probability models, as those commonly used in planning as inference, where the support of the used probability distributions directly corresponds to the states of percept, actions, thoughts. In summary, emergent generativity featured by the brain has these advantages: by default, the brain can learn the probability distributions of the hidden causes of any relevant cognitive element, the support (representation) of such distributions, and the probability dependencies between such causes. We speculate that the importance of these advantages might have contributed to lead evolution to endow the brain with spiking neurons rather than with firing-rate neurons (cf. [21, 37, 77]).

4 Conclusions

Goal-directed and planning processes can support flexible behaviour based on the use of general-purpose knowledge on the world. In recent years, it has been proposed that planning processes in the brain are based on probabilistic representations of the world and inferences on them. This proposal is very interesting but it encounters the great challenge of explaining how such representations and inferences might be grounded on the actual neural computations of the brain. Recently, some models have been proposed to ground some probability inference mechanisms, such as Hidden Markov Models and Partially Observable Markov Decision Processes, on the spiking stochastic events exhibited by the brain neurons and their connectivity patterns and plasticity mechanisms.

Here we propose a spiking neural-network architecture facing two important problems not solved by the state-of-the-art models bridging planning as inference and brain-like mechanisms, namely the problem of learning the world model contextually to its use for planning, and the problem of learning such world model in an autonomous fashion based on unsupervised learning processes. The architecture has been validated with data from human participants engaged in solving a visuomotor behavioural test that requires the discovery of the correct actions to associate to some stimuli [15]. The architecture has reproduced the target behaviour, has furnished an explanation of the mechanisms possibly underlying it, and has proposed predictions testable in future empirical experiments.

To overcome the two mentioned problems, the architecture proposes two novel mechanisms that the brain might use to solve them. First, it introduces a new arbitration mechanism that leads the model to plan and act to pursue the goal, or to explore to train the world model, on the basis of the knowledge of the world model itself: this knowledge is measured as the entropy of the goal-conditioned probability distribution of future states and actions expressed by the world model. Second, the model is able to autonomously learn the world model by integrating an STDP unsupervised learning rule proposed in the literature [49], with a world model based on a HMM whose observations involve not only world states but also actions, and using a goal representation to condition the probability distribution expressed by the world model.

We acknowledge that the model has various limitations that might be improved in future work. A first one concerns the passage from neurons firing at discrete times to neurons firing in continuous time. This might be done using the inhomogeneous Poisson process used in [49]. Although this would not change the theoretical contribution of the model, it might simplify a comparison of the model functioning with real data from the brain at a finer temporal level with respect to what done here.

A further issue to face would be to use other tasks with respect to the one considered here [15], for example to develop the model to consider tasks requiring longer sequences of states and actions as was done in [38, 39]. The latter works also suggest the interesting possibility of employing the model to control autonomous robots to test its robustness and capacity to scale-up to more complex tasks.

A relevant issue to face in future work concerns the new arbitration mechanism proposed with the model. The entropy measure at the core of such arbitration mechanism is grounded on the probability distribution of neurons. However, the mechanism using such information to arbitrate between planning and exploration is now hardwired. Future work should thus aim to implement this process with neural mechanisms. For example, the entropy measure might be 'read out' by an additional neural layer that could then selectively inhibit either the planning or the exploration component.

Another improvement of the model might involve the full development of a habitual component. Here we did not introduce such component as the target experiment covered a short learning time not allowing the formation of habits, so we focused on considering the exploration/exploitation processes involved in the early phases of learning of new tasks. Future work might however also consider the formation of habits, for example by targeting additional experiments involving long 'over-training' periods. This could be done with components analogous to the exploration component used here, but using slow reinforcement learning processes to represent the slow formation of habits favouring generalisation. The addition of habit learning processes would also require the introduction of a further arbitration mechanism as those proposed in [47, 48] to harmonise goal-directed and habitual behaviour.

A further possible improvement of the model concerns the treatment of goals. These are now selected externally and represented in a simple way. Goals could instead be represented in more realistic ways, for example through mechanisms mimicking working memory [78], and could be selected in autonomous ways, for example based on motivational mechanisms [72, 79].

A last possible improvement of the model concerns the possibility of testing and constraining the model not only at the behavioural level, as done here (and as also done by previous probabilistic models investigating arbitration mechanisms in goal-directed behaviour, e.g. [47, 48]) but also at the neural level, for example based on data collected on similar experiments [80, 81]. This might for example be done through techniques such as *Representational*

Similarity Analysis [82] that uses brain-imaging data to map the components of neural models to areas of the brain that possibly implement analogous functions.

Notwithstanding these limitations and possible improvements, we think the proposed architecture represents a further step towards the realisation of models that implement probabilistic versions of goal-directed processes on the basis of brain-like mechanisms, in particular spiking neurons, competitive circuits, and STDP unsupervised learning rules. In particular, the model contributes to formulate new hypotheses on how the brain might acquire the world model needed for planning in a fully autonomous way while at the same time using it for planning.

Acknowledgments

We thank Giovanni Pezzulo for suggesting links between our model and the literature on active inference. We also thank Andrea Mattera for feedback on the manuscript.

Author Contributions

Conceptualization: Ruggero Basanisi, Emilio Cartoni, Gianluca Baldassarre.

Data curation: Ruggero Basanisi.

Formal analysis: Ruggero Basanisi, Gianluca Baldassarre.

Funding acquisition: Andrea Brovelli, Gianluca Baldassarre.

Investigation: Ruggero Basanisi, Andrea Brovelli, Emilio Cartoni, Gianluca Baldassarre.

Methodology: Ruggero Basanisi, Emilio Cartoni, Gianluca Baldassarre.

Resources: Andrea Brovelli.

Software: Ruggero Basanisi, Emilio Cartoni.

Supervision: Andrea Brovelli, Emilio Cartoni, Gianluca Baldassarre.

Validation: Ruggero Basanisi, Andrea Brovelli, Emilio Cartoni, Gianluca Baldassarre.

Visualization: Ruggero Basanisi, Gianluca Baldassarre.

Writing – original draft: Ruggero Basanisi, Emilio Cartoni, Gianluca Baldassarre.

Writing – review & editing: Ruggero Basanisi, Emilio Cartoni, Gianluca Baldassarre.

References

1. Dickinson A, Balleine B. Motivational control of goal-directed action. Animal Learning & Behavior. 1994; 22(1):1–18.
2. Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998; 37(4):407–419. https://doi.org/10.1016/S0028-3908(98)00033-1 PMID: 9704982
3. Dolan R, Dayan P. Goals and Habits in the Brain. Neuron. 2013; 80(2):312–325. https://doi.org/10.1016/j.neuron.2013.09.007 PMID: 24139036
4. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge, MA: The MIT Press; 1998.
5. Sutton RS. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: Proceedings of the seventh international conference on machine learning. Vol. 216; 1990. p. 216–224.
6. Baldassarre G. Planning with neural networks and reinforcement learning [PhD Thesis]. Computer Science Department, University of Essex. Colchester, UK; 2002.
7. Baldassarre G. Forward and bidirectional planning based on reinforcement learning and neural networks in a simulated robot. In: Butz M, Sigaud O, Gérard P, editors. Anticipatory behaviour in adaptive

learning systems. Vol. 2684 of Lecture Notes in Artificial Intelligence. Berlin: Springer Verlag; 2003. p. 179–200.

8. Botvinick MM, Niv Y, Barto A. Hierarchically organized behavior and its neural foundations: A reinforcement-learning perspective. Cognition. 2008; 113(3):262–280. https://doi.org/10.1016/j.cognition.2008.08.011 PMID: 18926527
9. Balleine BW, Dezfouli A, Ito M, Doya K. Hierarchical control of goal-directed action in the cortical–basal ganglia network. Current Opinion in Behavioral Sciences. 2015; 5:1–7. https://doi.org/10.1016/j.cobeha.2015.06.001
10. Mannella F, Gurney K, Baldassarre G. The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. Frontiers in Behavioral Neuroscience. 2013; 7. https://doi.org/10.3389/fnbeh.2013.00135 PMID: 24167476
11. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. Upper Saddle River, NJ: Prentice Hall; 2003.
12. Steels L, Brooks R. The artificial life route to artificial intelligence: Building embodied, situated agents. Routledge; 2018.
13. Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, et al. A neural signature of hierarchical reinforcement learning. Neuron. 2011; 71(2):370–379. https://doi.org/10.1016/j.neuron.2011.05.042 PMID: 21791294
14. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. The role of the dorsomedial striatum in instrumental conditioning. Europearn Journal of Neuroscience. 2005; 22(2):513–523. https://doi.org/10.1111/j.1460-9568.2005.04218.x PMID: 16045504
15. Brovelli A, Laksiri N, Nazarian B, Meunier M, Boussaoud D. Understanding the Neural Computations of Arbitrary Visuomotor Learning through fMRI and Associative Learning Theory. Cerebral Cortex. 2008; 18(7):1485–1495. https://doi.org/10.1093/cercor/bhm198 PMID: 18033767
16. Brovelli A, Nazarian B, Meunier M, Boussaoud D. Differential roles of caudate nucleus and putamen during instrumental learning. NeuroImage. 2011; 57(4):1580–1590. https://doi.org/10.1016/j.neuroimage.2011.05.059 PMID: 21664278
17. Jahanshahi M, Obeso I, Rothwell JC, Obeso JA. A fronto–striato–subthalamic–pallidal network for goal-directed and habitual inhibition. Nature Reviews Neuroscience. 2015; 16(12):719–732. https://doi.org/10.1038/nrn4038 PMID: 26530468
18. Caligiore D, Arbib MA, Miall CR, Baldassarre G. The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia. Neuroscience and Biobehavioral Reviews. 2019; 100:19–34. https://doi.org/10.1016/j.neubiorev.2019.02.008 PMID: 30790636
19. Helmholtz H. Concerning the perceptions in general. In: Southall J, editor. Treatise on physiological optics (3rd ed., Vol. III, Translation 1962). New York: Dover; 1866. p. 214–230.
20. Dayan P, Hinton GE, Neal RM, Zemel RS. The Helmholtz machine. Neural computation. 1995; 7(5):889–904. https://doi.org/10.1162/neco.1995.7.5.889 PMID: 7584891
21. Doya K, Ishii S, Pouget A, Rao RPN, editors. The Bayesian Brain: Probabilistic Approaches to Neural Coding. Cambridge, MA: MIT Press; 2007.
22. Friston K. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience. 2010; 11(2):127–138. https://doi.org/10.1038/nrn2787 PMID: 20068583
23. Griffiths TL, Kemp C, Tenenbaum JB. Bayesian models of cognition. Cambridge, UK: Cambridge University Press; 2008.
24. Toussaint M, Storkey A. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In: Proceedings of the 23rd international conference on Machine learning. ACM; 2006. p. 945–952.
25. Botvinick M, Toussaint M. Planning as inference. Trends in Cognitive Sciences. 2012; 16(10):485–488. https://doi.org/10.1016/j.tics.2012.08.006 PMID: 22940577
26. Kappen HJ, Gómez V, Opper M. Optimal control as a graphical model inference problem. Machine learning. 2012; 87(2):159–182. https://doi.org/10.1007/s10994-012-5278-7
27. Rao RP, Olshausen BA, Lewicki MS. Probabilistic models of the brain: Perception and neural function. Boston, MA: MIT press; 2002.
28. Jones M, Love BC. Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences. 2011; 34(4):169–88; disuccsion 188–231. https://doi.org/10.1017/S0140525X10003134 PMID: 21864419
29. Sharma S, Voelker A, Eliasmith C. A Spiking Neural Bayesian Model of Life Span Inference. In: CogSci; 2017. p. 3131–3136.

30. Maass W. Networks of spiking neurons: the third generation of neural network models. Neural networks. 1997; 10(9):1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7

31. Deneve S. Bayesian inference in spiking neurons. In: Advances in neural information processing systems; 2005. p. 353–360.

32. Buesing L, Bill J, Nessler B, Maass W. Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons. PLoS Computational Biology. 2011; 7(11):e1002211. https://doi.org/10.1371/journal.pcbi.1002211 PMID: 22096452

33. Orhan AE, Ma WJ. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. Nature communications. 2017; 8:138. https://doi.org/10.1038/s41467-017-00181-8 PMID: 28743932

34. Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nature Neuroscience. 2013; 16(9):1170–1178. https://doi.org/10.1038/nn.3495 PMID: 23955561

35. Del Moral P. Nonlinear filtering: Interacting particle resolution. Markov Processes and Related Fields. 1996; 2(4):555–580.

36. Wang X, Li T, Sun S, Corchado JM. A survey of recent advances in particle filters and remaining challenges for multitarget tracking. Sensors. 2017; 17(12):2707. https://doi.org/10.3390/s17122707 PMID: 29168772

37. Huang Y, Rao RP. Neurons as Monte Carlo Samplers: Bayesian Inference and Learning in Spiking Networks. In: Advances in neural information processing systems; 2014. p. 1943–1951.

38. Rueckert E, Kappel D, Tanneberg D, Pecevski D, Peters J. Recurrent Spiking Networks Solve Planning Tasks. Scientific Reports. 2016; 6(1). https://doi.org/10.1038/srep21142 PMID: 26888174

39. Tanneberg D, Paraschos A, Peters J, Rueckert E. Deep spiking networks for model-based planning in humanoids. In: Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on. IEEE; 2016. p. 656–661. Available from: http://ieeexplore.ieee.org/abstract/document/7803344/.

40. Friedrich J, Lengyel M. Goal-Directed Decision Making with Spiking Neurons. Journal of Neuroscience. 2016; 36(5):1529–1546. https://doi.org/10.1523/JNEUROSCI.2854-15.2016 PMID: 26843636

41. Solway A, Botvinick MM. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychological Review. 2012; 119(1):120–154. https://doi.org/10.1037/a0026435 PMID: 22229491

42. Rückert EA, Neumann G, Toussaint M, Maass W. Learned graphical models for probabilistic planning provide a new class of movement primitives. Frontiers in Computational Neuroscience. 2013; 6. https://doi.org/10.3389/fncom.2012.00097 PMID: 23293598

43. Passingham RE, Wise SP. The neurobiology of the prefrontal cortex: anatomy, evolution, and the origin of insight. Vol. 50. Oxford: Oxford University Press; 2012.

44. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441(7095):876–879. https://doi.org/10.1038/nature04766 PMID: 16778890

45. Mehlhorn K, Newell BR, Todd PM, Lee MD, Morgan K, Braithwaite VA, et al. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision. 2015; 2(3):191–215. https://doi.org/10.1037/dec0000033

46. Tschantz A, Seth AK, Buckley CL. Learning action-oriented models through active inference. PLoS computational biology. 2020; 16:e1007805. https://doi.org/10.1371/journal.pcbi.1007805 PMID: 32324758

47. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience. 2005; 8(12):1704–1711. https://doi.org/10.1038/nn1560 PMID: 16286932

48. Viejo G, Khamassi M, Brovelli A, Girard B. Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. Frontiers in Behavioral Neuroscience. 2015; 9. https://doi.org/10.3389/fnbeh.2015.00225 PMID: 26379518

49. Kappel D, Nessler B, Maass W. STDP Installs in Winner-Take-All Circuits an Online Approximation to Hidden Markov Model Learning. PLoS Computational Biology. 2014; 10(3):e1003511. https://doi.org/10.1371/journal.pcbi.1003511 PMID: 24675787

50. Luppino G, Rizzolatti G. The Organization of the Frontal Motor Cortex. News in physiological sciences. 2000; 15:219–224. PMID: 11390914

51. Thill S, Caligiore D, Borghi AM, Ziemke T, Baldassarre G. Theories and computational models of affordance and mirror systems: An integrative review. Neuroscience and Biobehavioral Reviews. 2013; 37:491–521. https://doi.org/10.1016/j.neubiorev.2013.01.012 PMID: 23333761

52. Treves A, Rolls ET. Computational analysis of the role of the hippocampus in memory. Hippocampus. 1994; 4(3):374–391. https://doi.org/10.1002/hipo.450040319 PMID: 7842058

53. Basanisi R, Brovelli A, Cartoni E, Baldassarre G. A spiking neural-network model of goal-directed behaviour. bioRxiv. 2019; https://doi.org/10.1101/867366.

54. Klein RM. Inhibition of return. Trends in Cognitive Sciences. 2000; 4(4):138–147. https://doi.org/10.1016/S1364-6613(00)01452-2 PMID: 10740278

55. Bishop CM. Pattern recognition and machine learning. springer; 2006.

56. Neal RM, Hinton GE. A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Learning in graphical models. Springer; 1998. p. 355–368.

57. Bishop CM. Pattern recognition and machine learning. Berlin: Springer; 2006.

58. Jolivet R, Rauch A, Lüscher HR, Gerstner W. Predicting spike timing of neocortical pyramidal neurons by simple threshold models. Journal of computational neuroscience. 2006; 21(1):35–49. https://doi.org/10.1007/s10827-006-7074-5 PMID: 16633938

59. Dan Y, Poo Mm. Spike timing-dependent plasticity of neural circuits. Neuron. 2004; 44(1):23–30. https://doi.org/10.1016/j.neuron.2004.09.007 PMID: 15450157

60. Feldman D. The Spike-Timing Dependence of Plasticity. Neuron. 2012; 75(4):556–571. https://doi.org/10.1016/j.neuron.2012.08.001 PMID: 22920249

61. Markram H, Gerstner W, Sjöström PJ. Spike-Timing-Dependent Plasticity: A Comprehensive Overview. Frontiers in Synaptic Neuroscience. 2012; 4. https://doi.org/10.3389/fnsyn.2012.00002 PMID: 22807913

62. Zappacosta S, Mannella F, Mirolli M, Baldassarre G. General differential Hebbian learning: Capturing temporal relations between events in neural networks and the brain. Plos Computational Biology. 2018; 14(8):e1006227. https://doi.org/10.1371/journal.pcbi.1006227 PMID: 30153263

63. Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer; 2001.

64. Miikkulainen R, Bednar JA, Choe Y, Sirosh J. Computational maps in the visual cortex. Springer; 2006.

65. Gibson JJ. The Ecological Approach to Visual Perception. Boston, MA: Houghton Mifflin; 1979.

66. Baldassarre G, Lord W, Granato G, Santucci VG. An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning. Frontiers in Neurorobotics. publ2019; 13(45). https://doi.org/10.3389/fnbot.2019.00045 PMID: 31402859

67. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems. MIT Press; 2014. p. 2672–2680.

68. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv preprint arXiv:13126114. 2013.

69. Gregor K, Papamakarios G, Besse F, Buesing L, Weber T. Temporal Difference Variational Auto-Encoder. arXiv preprint arXiv:180603107. 2018.

70. Hinton GE. Training products of experts by minimizing contrastive divergence. Neural computation. 1995; 14(8):1771–1800. https://doi.org/10.1162/089976602760128018

71. Goodfellow I, Bengio Y, Courville A. Deep Learning. Boston, MA: The MIT Press; 2017.

72. Granato G, Baldassarre G. Human Flexible Goal-directed Behavior and the Manipulation of Internal Representations: A Computational Model. PsyArXiv. 2019; p. e1–33.

73. Newman MEJ. Complex Systems: A Survey. arXiv preprint arXiv:11121440. 2011; 79:800–810.

74. Maass W. On the computational power of winner-take-all. Neural computation. 2000; 12(11):2519–2535. https://doi.org/10.1162/089976600300014827 PMID: 11110125

75. Nessler B, Pfeiffer M, Buesing L, Maass W. Bayesian Computation Emerges in Generic Cortical Microcircuits through Spike-Timing-Dependent Plasticity. PLoS Computational Biology. 2013; 9(4):e1003037. https://doi.org/10.1371/journal.pcbi.1003037 PMID: 23633941

76. Bill J, Buesing L, Habenschuss S, Nessler B, Maass W, Legenstein R. Distributed Bayesian Computation and Self-Organized Learning in Sheets of Spiking Neurons with Local Lateral Inhibition. PLOS ONE. 2015; 10(8):e0134356. https://doi.org/10.1371/journal.pone.0134356 PMID: 26284370

77. Zheng Y, Jia S, Yu Z, Huang T, Liu JK, Tian Y. Probabilistic inference of binary Markov random fields in spiking neural networks through mean-field approximation. Neural networks. 2020; 126:42–51. https://doi.org/10.1016/j.neunet.2020.03.003 PMID: 32197212

78. O'Reilly RC, Frank MJ. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Computation. 2006; 18(2):283–328. https://doi.org/10.1162/089976606775093909 PMID: 16378516

79. Mannella F, Mirolli M, Baldassarre G. Goal-Directed Behavior and Instrumental Devaluation: A Neural System-Level Computational Model. Frontiers in Behavioral Neuroscience. 2016; 10(181):e1–27. https://doi.org/10.3389/fnbeh.2016.00181 PMID: 27803652

80. Brovelli A, Chicharro D, Badier JM, Wang H, Jirsa V. Characterization of Cortical Networks and Cortico-cortical Functional Connectivity Mediating Arbitrary Visuomotor Mapping. Journal of Neuroscience. 2015; 35(37):12643–12658. https://doi.org/10.1523/JNEUROSCI.4892-14.2015 PMID: 26377456

81. Brovelli A, Badier JM, Bonini F, Bartolomei F, Coulon O, Auzias G. Dynamic reconfiguration of visuomotor-related functional connectivity networks. Journal of Neuroscience. 2017; 37(4):839–853. https://doi.org/10.1523/JNEUROSCI.1672-16.2016 PMID: 28123020

82. Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis—Connecting the branches of systems neuroscience. Frontiers in systems neuroscience. 2008; 2:4. https://doi.org/10.3389/neuro.06.004.2008 PMID: 19104670

Section 3. Beta oscillations in the monkey striatum encodes reward prediction error

Beta oscillations in the monkey striatum encodes reward prediction error

Basanisi, R.[1] , Marche, K.[1, 2] , Combrisson, E.[1], Apicella, P.[1 §], Brovelli, A.[1 §]

1 Institut de Neurosciences de la Timone, Aix Marseille Université, UMR 7289 CNRS, 13005, Marseille, France
2 Wellcome Centre for Integrative Neuroimaging, Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom.

§ P.A. and A.B are co-senior authors

Corresponding authors:
Ruggero Basanisi
ruggero.basanisi@gmail.com
Institut de Neurosciences de la Timone (INT),
UMR 7289 CNRS, Aix Marseille University,
Campus de Santé Timone,
27 Bd. Jean Moulin,
13385 Marseille, France

Paul Apicella
paul.apicella@univ-amu.fr
Institut de Neurosciences de la Timone (INT),
UMR 7289 CNRS, Aix Marseille University,
Campus de Santé Timone,
27 Bd. Jean Moulin,
13385 Marseille, France

Andrea Brovelli
andrea.brovelli@univ-amu.fr
Institut de Neurosciences de la Timone (INT),
UMR 7289 CNRS, Aix Marseille University,
Campus de Santé Timone,
27 Bd. Jean Moulin,
13385 Marseille, France

Keywords
Basal ganglia, local field potentials, choice behavior, learning, reward prediction error

Abstract

Reward prediction errors (RPEs) reflect the difference between obtained and predicted rewards, and they are a building block of basic forms of reinforcement learning. RPE signals are encoded by the activity of midbrain dopaminergic neurons that innervate the striatum and frontal cortex, suggesting that RPE signals are integrated in cortico-basal ganglia circuits. In the current study, we investigated the participation of the different territories of the striatum in the encoding of RPE. To do so, we recorded local field potentials (LFPs) in the striatum of two rhesus monkeys performing a task involving a choice among options for movement with different reward probabilities. The trial-by-trial evolution of RPE was estimated using a reinforcement learning model fitted on monkeys' choice behavior. We found that changes in beta band oscillations (15-30 Hz) during the outcome period appear consistent with RPE encoding. Moreover, the learning-relevant outcome information contained in beta oscillations increased along a dorsolateral-to-ventromedial gradient. These region-specific changes in LFP activity suggest a relationship between beta oscillations in the striatum and the evaluation of outcome based on reward feedback, highlighting a specific contribution of the ventral striatum to the updating of choice behavior.

Introduction

The striatum is the major component of the basal ganglia and it plays a key role in action selection and reward-guided learning under the influence of ascending dopaminergic projections from the ventral midbrain. Previous research in monkey neurophysiology (Samejima et al., 2005; Lau and Glimcher, 2007; Seo et al., 2012; Yamada et al., 2013) and functional magnetic resonance imaging (fMRI) in humans (Balleine et al., 2007; Delgado et al., 2005; Wang et al., 2016) has identified neural signals coding action-value in the striatum. Striatal neuronal activity has also been reported to reflect the difference between received and expected rewards, the so-called reward prediction error or RPE (Sutton and Barto, 1998). RPE signals are thought to be crucial for the update of action values (Schultz, 2007; Fujiyama et al., 2015; Schultz, 2016a, 2016b). Several studies have shown evidence, in the striatum of both monkeys and rodents, that output neurons (Roesch et al., 2009; Oyama et al., 2010; Asaad and Eskandar, 2011) and putative interneurons (Apicella et al., 2009; Stalnaker et al., 2012) encode RPE to promote reward-guided learning. fMRI studies in humans have assessed the role of striatum, in particular its ventral part, in encoding RPE (O'Doherty, 2004; O'Doherty et al., 2007; Bray and O'Doherty, 2007; Park et al., 2012; Kumar et al., 2018; Calderon et al., 2021). Another fMRI study proposed that RPEs deriving from different types of reward can recruit distinct partially overlapping striatal circuits (Valentin and O'Doherty, 2009). Despite these findings highlighting the involvement of striatum in RPE encoding, less is still known about neurophysiological activity supporting RPE learning across striatal regions.

Among neural signals that may serve as potential physiological markers for the processing of information in basal ganglia circuits, there is a strong emphasis on local field potential (LFPs) that are supposed to reflect the synchronous activity of populations of neurons in a given brain region (Goldberg, 2004; Brown and Williams, 2005). In particular, oscillations in the beta-frequency band (typically about 15–30 Hz) have been related to motor function. Indeed, increases in beta LFP oscillatory activity have been linked to motor impairments in patients with Parkinson's disease (Brown, 2007; Jenkinson and Brown, 2011) and animals with experimentally induced Parkinson-like states (Wichmann et al., 1994; Nini et al.,

1995; Deffains et al., 2016; Kondabolu et al., 2016). This beta LFP oscillatory activity has been detected at different levels of the basal ganglia network, including the striatum. Besides their well known link to pathology, striatal beta oscillations were also present in normal behaving rats (Berke et al., 2004; Leventhal et al., 2012; Schmidt et al., 2013) and monkeys (Courtemanche et al., 2003; Bartolo et al., 2014). Numerous studies have provided evidence that this oscillatory activity can be modulated during specific phases of behavioral tasks, possibly reflecting a wide range of cognitive processes, such as modulation of task performances through reinforcement learning (Feingold et al., 2015), response to attentional cues (Banaie Boroujeni et al., 2020), and cues utilization for action programming (Leventhal et al., 2012). Prior studies have shown that striatal beta activity is modulated by reward delivered on correct trials during learning tasks in rats (Howe et al., 2011) and by different task parameters, including reward value, in monkeys (Schwerdt et al., 2020). So far, no consensus has been achieved on the functional implications of such changes. In particular, it is not clear whether RPE signals during the processing of action outcomes may influence striatal beta activity.

Moreover, previous research has pointed out that striatal beta oscillations and their relation to motor and reward processing may occur in a regionally dependent manner (Howe et al., 2011; Schwerdt et al., 2020). It has long been recognized that the striatum is divided into three functional domains (i.e., motor, associative and limbic) (Parent, 1990; Lanciego et al., 2012). In primates, the motor division is located in posterior dorsolateral portions of the putamen, while the associative and limbic divisions encompass dorsal and ventral portions of the anterior caudate nucleus and putamen, respectively (Nakano et al., 2000; Liljeholm and O'Doherty, 2012; Eisinger et al., 2018). Several fMRI studies in humans have reported that the processing of reward-related information, including RPE, is dominant in the ventral rather than the dorsal striatum (Apicella et al., 1991; O'Doherty, 2004; Abler et al., 2006; O'Doherty et al., 2007; Hare et al., 2008). In addition, neuronal recordings in rats have shown that the nucleus accumbens is important for updating choice behaviors (Ito and Doya, 2009). To our knowledge, no experiment in the monkey has provided an in-depth analysis comparing changes in beta oscillations across distinct territories of the striatum in relation to RPE signals.

In the present study, we test the hypothesis that the striatum encodes RPE signals according to an anatomo-functional gradient. To do so, we studied LFP activity recorded at different sites in the striatum of two monkeys performing a free-choice probabilistic learning task. The aim was to characterize the relationship between beta oscillations and choice behavior and its possible role in encoding RPE. The results indicate that changes in striatal beta-band activity play a role in encoding RPEs along an anatomo-functional gradient, which shows a dominant component in the ventral, rather than the dorsal striatum.

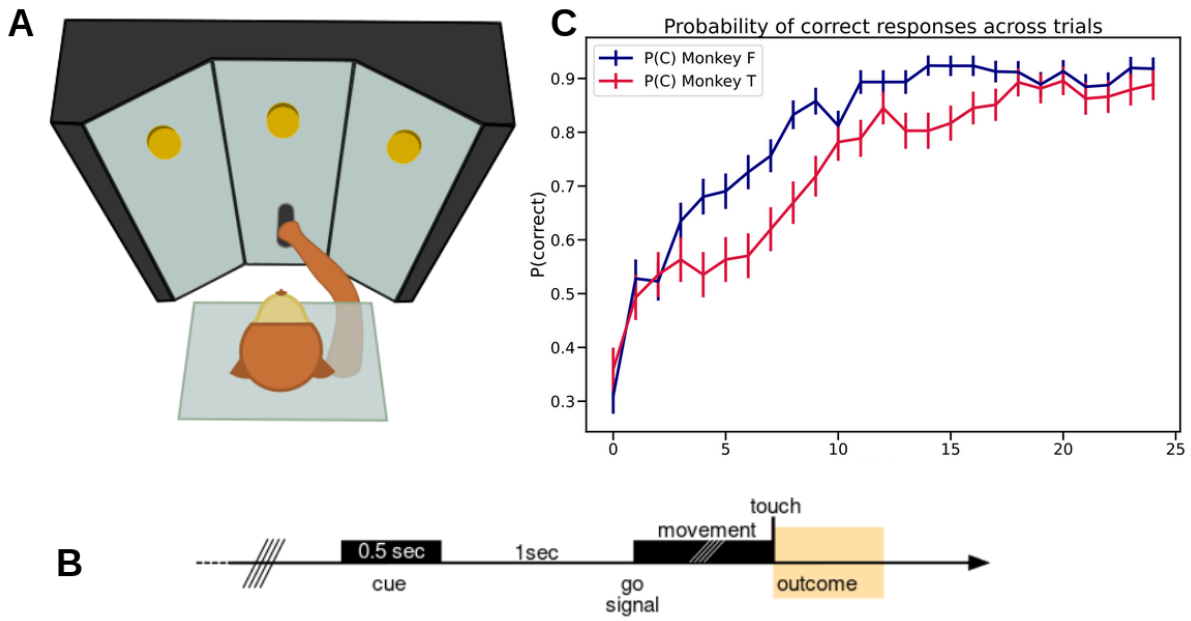


Figure 1. *Sequence of events and performance in the choice task.*
A) Experimental set-up of the free-choice probabilistic learning task. The monkey sat in a cage with three buttons in front. Keeping the hand on a metal bar allowed the trial to start.
B) Single trial time course. After the beginning of a new trial, a cue signal warned the monkey about the arrival of the go signal in one second. When the go signal appeared, the monkey performed the motor response towards one of the three yellow targets (no cue about the most rewarding response was given). At target touch, the monkey received feedback (reward or no reward). Correlates of the RPE signals were analysed in the time window indicated as "outcome".
C) The two curves depict monkeys' performances as the probability of correct response averaged across learning sessions for monkey F (blue) and T (red).

Results

In this work, we studied whether striatal beta-band (15-30Hz) oscillations are involved in encoding RPE in different striatal territories. To do so, we recorded local field potentials (LFPs) from the striatum of two macaque monkeys while performing a free-choice probabilistic learning task.

The analysis of behavioral performances during task execution confirmed that both monkeys learned by trial-and-error over the course of each session which target was most rewarding. Each session was characterised by an initial exploration phase that allowed monkeys to find the most rewarding action, followed by a phase in which monkeys preferentially chose the most rewarding target until the end of the block. In order to quantify behavioral performance across monkeys, we aligned all the sessions to the beginning of each block and computed the probability of correct response across trials. The probability of correct response quantified the monkey's ability to choose the most rewarding button among the three options. As we can see from the progression of the curves in **Figure 1**, 15-20 trials were sufficient for both monkeys to figure out the position of the most rewarding target.

LPF power was analysed using a reinforcement learning model-based approach. Finally, we used information theory tools and cluster based statistics together with linear regression models to perform statistical analyses.

Reward modulates beta band power

We then investigated whether modulations in striatal beta-band activity differed among rewarded and unrewarded trials. To do so, we collected all the single trial time-frequency maps for each condition (rewarded and unrewarded). Then we performed a two-sided t-test analysis across the two obtained data samples, and then we Bonferroni corrected the p-values on the total number of points in the time-frequency matrix. The significant clusters ($p<0.05$, **Figure 2**) were observed for both monkeys in the beta band. On the other hand, the clusters were centred around 25 Hz for monkey F and around 30 Hz for monkey T. We used those two central

frequencies to perform subject-specific analyses of beta band power using the multitaper method.

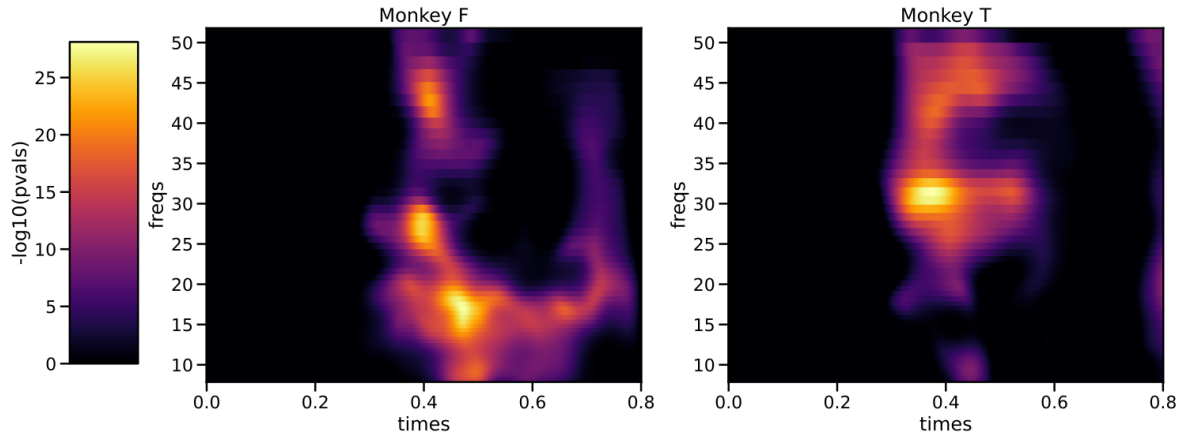


Figure 2. *Statistical power of the t-test performed on the time-frequency power map for each monkey, when contrasting rewarded vs. non-rewarded trials. Time 0 corresponds to button press. The p-values were Bonferroni corrected across the total number of points in the map. To simplify visualization, we are showing the -log10 corresponding value (-log10(0.05) = ~1.3).*

In order to study whether such modulation in beta band power reflected reward prediction errors (RPEs), we fitted a standard Q-learning model to the single-session behavioral data.

From the model, we extracted two values, the RPE and its absolute value, and we used these two model-based variables together with three other model-free variables (reaction times, movement times and chosen action) to fit a multiple linear regression model with respect to the beta band neurophysiological data. Then we used the obtained distributions of angular coefficients to compute a two tailed t-test. As you can see in **Table 1**, the only significant regressor related to the examined period of activity was the RPE.

Thus, in order to identify neural correlates of RPEs, we then computed the mutual information between evolution of RPE and beta band activity across trials in a time-resolved manner. Statistical analysis was performed using cluster-based statistics combined with permutation tests.

As shown in **Figure 3**, we found a significant relation, quantified by means of MI, between RPEs and beta band activity. In both monkeys, the time-course of MI

increased around 200 msec, peaked around 450 msec after outcome onset and lasted a total of approximately 550msec. Significant values ($p<0.05$) are represented in the plot by the continuous line; this measure can be interpreted as a trial by trial covariation in time of the electrophysiological and the behavioral measures.This result is confirmed by cluster-based statistics and permutation tests. The statistical framework that we used is detailed in the **Materials and Methods** section. This result confirms that beta band power variation in the striatum is differentially modulated by feedback type (i.e., presence or absence of rewards) and encodes RPE signals.

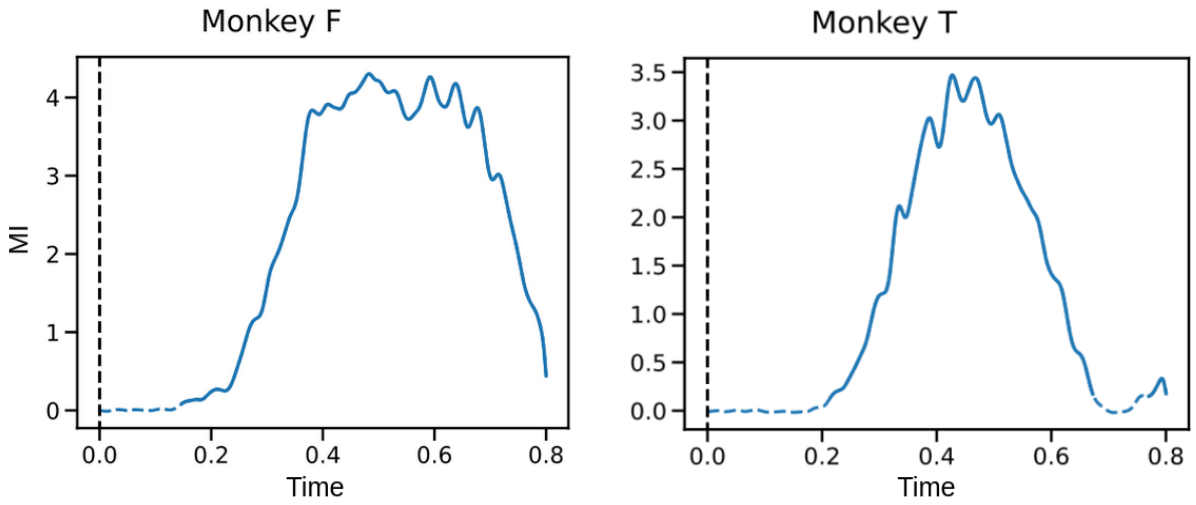


Figure 3. *Mutual Information (MI) between beta-band oscillation and RPE. The dashed lines represent non significant values ($p \geq 0.05$), while the continuous lines represent significant values ($p < 0.05$).*

Information about RPE dissociates striatal regions

We next investigated whether the encoding of RPEs by beta-band LPF power modulations differentially recruited the sensorimotor, associative and limbic territories of the striatum. **Figure 4** illustrates the spatial distribution of striatal sites at which we recorded LFPs in one monkey, as verified by histological analysis. We subdivided the recording sessions into different groups according to their spatial location in the striatum. In order to group recording sessions into homogeneous clusters, we used the KMeans algorithm applied to the 3-dimensional spatial coordinates (AP, ML and depth) of the recording sites. We set the number of clusters equal to six for each

territory (putament, caudate and nucleus accumbens), in order to retain a sufficient number of trials per cluster. Thus, we obtained a total of eighteen spatial clusters, as represented in **Figure 5A**. Once we obtained the clusters, we computed the MI between the RPEs and beta power as described in the previous section. We observed that the amount of information carried by the beta-band LFP power about the RPE is higher in the limbic striatum then slowly decreases in the associative territory to finally drop down in the motor striatum. In **Figure 5B**, each line corresponds to a striatal territory, and for each line the figures are ordered by the maximum value of the sum of the MI of each cluster. As in **Figure 4**, dashed lines correspond to non-significant time intervals, while full lines correspond to significant temporal clusters. As shown in **Figure 5B**, the number of significant clusters decreases across territories following this pattern.

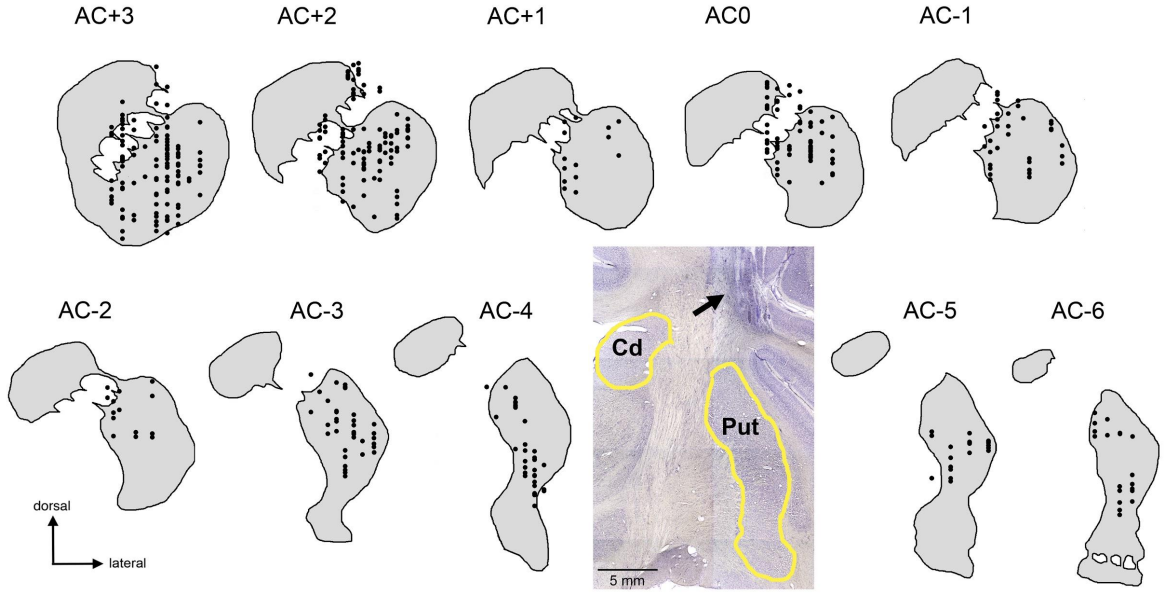


Figure 4. *Positions of all striatal recording sites in monkey F. Each dot corresponds to a single LFP recording site. Coronal sections are labeled in rostrocaudal stereotaxic planes according to distances from the anterior commissure (AC) used as a reference landmark.The inset shows a photomicrograph of a coronal section stained with Cresyl violet at the level of the posterior putamen (i.e., motor striatum) with visible traces of electrode tracks above the putamen. Cd, caudate nucleus; Put, putamen.*

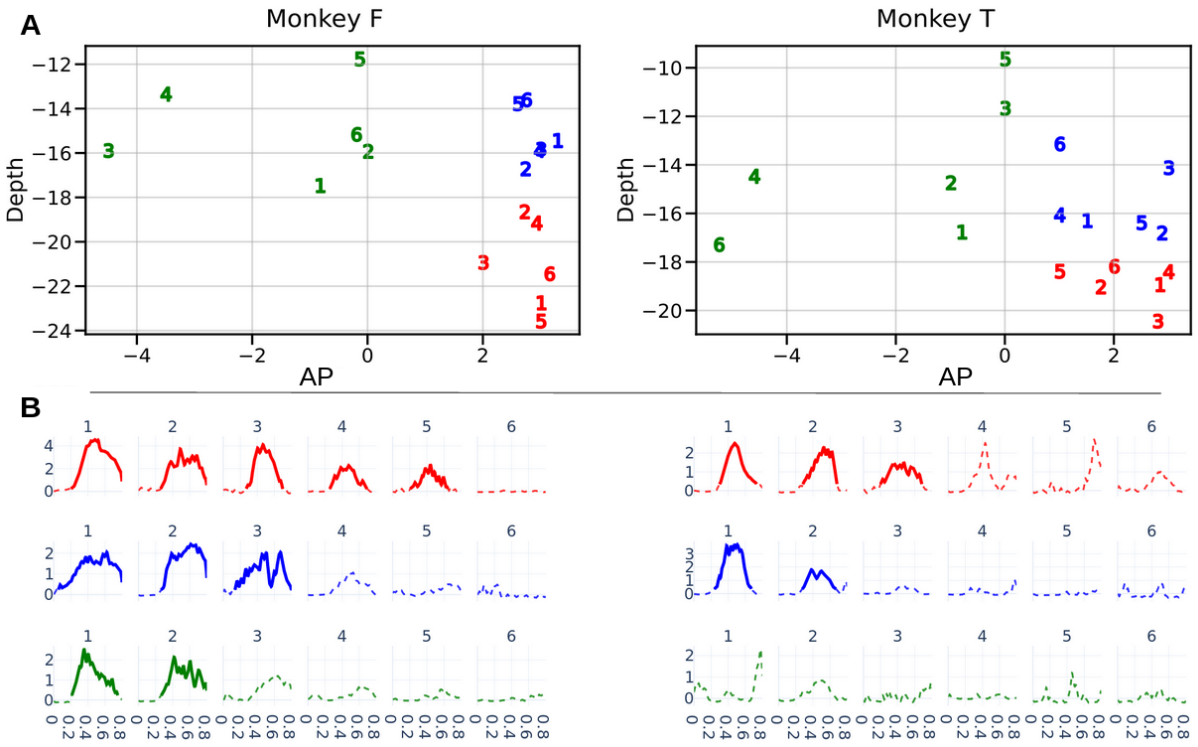


Figure 5. *A) two-dimensional spatial positions of the clusters of recording sites, for monkey F and T. Clusters are represented along their antero-posterior (AP) position (antero = positive numbers, posterior = negative numbers) and depth (deeper = lower number) of the recording site. Each color corresponds to an anatomo-functional region: red = limbic striatum, blue = associative striatum, green = motor striatum.*
B) MI computed in each of the clusters, for monkey F and T. The colour and the number associated with each cluster corresponds with the image on the top. From this image emerges how the number of clusters with a statistical significant increase of MI is higher in limbic striatum, to then progressively diminish in associative and motor striatum.

RPE follows a rostro-caudal and dorso-ventral gradient

We then assessed how the average amount of information about the RPE is distributed across striatum. To answer this question, we defined a rostro-caudal and dorso-ventral axis by taking the highest and the most posterior among electrodes' positions to define a referential point in space for each of the two monkeys. We computed the euclidean distance from the reference point to the center of each cluster, which allowed us to investigate the presence of linear or nonlinear relations between clusters' positions and functional effects (MI values). **Figure 6** shows an increase in RPE information with the distance from the referential point, toward the

rostral-ventral striatum. Linear correlation analysis revealed a significant and positive correlation (p-values < 0.05) for both monkeys. In other words, this result indicates that the amount of information about RPE signals follows an anatomical gradient, showing higher values in the rostro-ventral part of the striatum and gradual decrease towards the most dorso-caudal part.

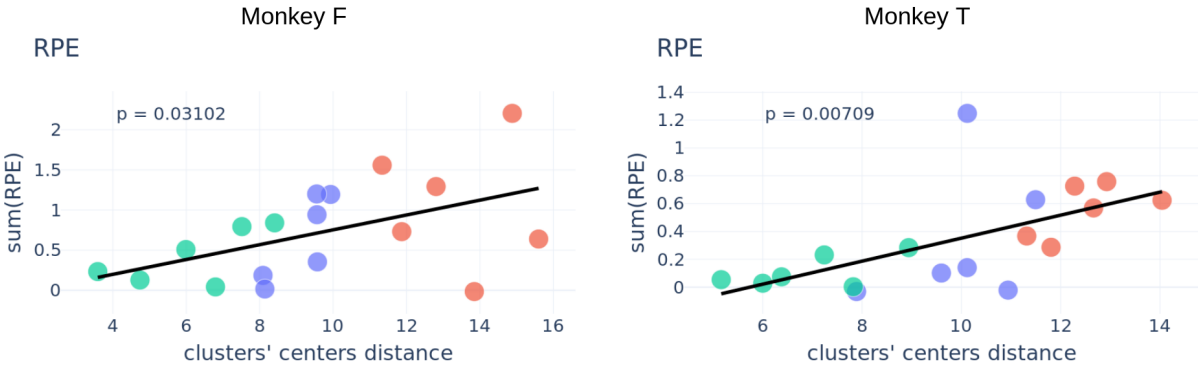


Figure 6. *Striatal gradient of the total RPE-beta band MI. On the y-axis we plotted the sum over the outcome time of the MI computed among RPE and beta band activity. On x-axis we plotted the distance of clusters from a reference point computed taking the AP coordinate of the most posterior recording site and the Depth coordinate of the higher recording site of each monkey. The linear regression with the associated p-value shown in the figure suggests an increasing gradient of RPE related activity toward the most rostro-ventral part of the striatum.*

Material and Methods

Experimental procedure and data acquisition

Experimental set-up and behavioral task

Two male adult rhesus monkeys (*Macaca mulatta*) were trained in an instrumental free-choice probabilistic learning task, in which they learned to choose among three options depending on the relative difference in reward probability associated with each option. All procedures were approved by the Institut de Neurosciences de la Timone Ethics Committee (Protocol A2-10-12) and were in accordance with guidelines from the National Institute of Health. Briefly, the surgically implanted monkeys were head-restrained to allow for stable electrophysiological recordings in different regions of the striatum.

The task required monkeys to choose among three spatial cues that were associated with different probabilities of liquid reward. Both monkeys were previously involved in other experiments studying single-neuron activity in the striatum during performance of simplified versions of the reaching task (Marche et al., 2017; Marche and Apicella, 2021). As shown in **Figure 1**, the experimental setup consisted of three targets (10-mm diameter) aligned horizontally (left, center, right), at the monkey's eye level, in a panel that was placed at a distance of 30 cm in front of the animal. The distance between targets was 10 cm. A two-color (red and green) light-emitting diode (LED) was located in the bottom of each target. Monkeys were trained to keep their hands on a metal rod, located on the lower part of the panel, at their waist level, as a starting position for the movement. A tube positioned directly in front of the animal's mouth dispensed small amounts of fruit juice (0.3 ml) as reinforcement.

Each trial was initiated when the monkey kept its hand on the rod for 1 s, after which all three LEDs were lit with a green color for 500 ms (cue onset). A fixed delay period of 1 s followed the offset of the cue. After the delay period, all three LEDs turned to red and this instructed the monkey to start a movement toward the chosen target. Once a target was touched, all three stimuli turned off and the monkey immediately received the associated outcome (reward or no reward) according to the

programmed schedule. Regardless of the outcome, the monkey had to bring the hand back on the rod to initiate a new trial. A new trial could not begin until the total duration of the current trial (6 s) had elapsed. Trials in which the monkey released the bar before trigger onset were aborted. If the monkey did not release the bar within a maximum time of 1 s after trigger onset or did not contact a target within a maximum time of 1 s after bar release, this was considered as incorrect.We tested monkeys in two learning contexts in which the probability of reward associated with each target was varied, a first ‘easy’ condition and a second ‘hard’ condition with relative reward probabilities of 70%-15%-15% and 50%-25%-25%.

Each condition was predetermined at the beginning of each block of trials and was changed from block to block. No explicit information regarding reward probabilities was available. Therefore, monkeys learned by trial-and-error the location of the most-rewarding target (i.e., the option with higher reward probability). The location of the best rewarded target was chosen pseudorandomly across trial blocks. There was no explicit information indicating transitions between blocks of trials and there was a varying number of trials per block (30-80 trials) to prevent anticipation of a block transition by the number of trials.

For each trial, we computed the reaction time (RT, defined as the time interval between the go signal and the bar release) and the movement time (MT, from the beginning of the movement to the target contact).

Acquisition of neurophysiological data

We used conventional techniques for recording single neuron activity from striatum (Marche et al., 2017). Monkeys were implanted with a recording chamber targeting the striatum, centered on the anterior commissure. This location allowed vertical access with custom-made glass-coated tungsten microelectrodes (impedance: 1–2.5 MΩ) to the putamen and caudate nucleus. Recordings were made in striatal sites where single-neuron activity was found, and the sites changed from session to session. LFPs from electrode were amplified (x 5000), bandpass filtered (3-150 Hz), and then sampled at 16.6 kHz by using a Power1401 Analog-Digital converter and a multi-channel acquisition software (Spike2, version 7.2; Cambridge Electronic Design).

Histological reconstructions

Recording sites were histologically verified in both animals, using several small electrolytic lesion marks in the putamen anterior and posterior to the anterior commissure (Marche et al., 2017). Upon completion of electrophysiological recordings, monkeys were anesthetized by using pentobarbital and perfused with 4% paraformaldehyde. Coronal brain slices (40 µm thickness) containing the striatum were prepared and stained by using Cresyl violet to identify the lesion marks. Electrode penetrations were reconstructed in serial sections through the striatum in each monkey.

Behavioral model

In order to model behavioral choices and estimate the evolution of RPEs during learning, we used a Q-learning model (Watkins and Dayan, 1992) from reinforcement learning theory (Sutton and Barto, 1998). Briefly, the Q-learning model updates action values through the Rescorla-Wagner learning rule (1972) expressed by the following equation:

$$Q_a(t+1) = Q_a(t) + \Delta Q \quad (1)$$

where $Q_a(t)$ corresponds to the value of action a=1, 2, 3 (three possible movements to 3 targets) at trial *t*, and ΔQ corresponds to the update value, also called Reward Prediction Error (RPE):

$$\Delta Q = RPE = \alpha \cdot (r(t) - Q_a(t)) \quad (2)$$

where α is the learning rate (which varies from 0 to 1) and r models the type of outcome. The r parameter takes values equal to 1 for a correct response, 0 if incorrect. Action values are then transformed into probabilities according to the softmax equation:

$$P_a(t) = exp(\beta Q_a(t)) \,/\, \sum_a exp(\beta Q_a(t)) \quad (3)$$

The coefficient β is termed the inverse 'temperature': low β (less than 1) causes all actions to be (nearly) equiprobable, whereas high β (greater than 1) amplifies the differences in association values.

We identified the set of parameters that best fitted the behavioural data using a maximum likelihood approach. The model was fitted separately for each block of trials and learning session.

For each learning session, we varied the learning rate λ from 0.1 to 1 (in steps of 0.01) and β was varied from 1 to 10 (in steps of 0.2). The two free variables of the model that we fitted are the learning rate of the learning rule (λ) and the inverse of the temperature used by the softmax function, and we used a grid search algorithm to find the best fitting couple of values. For each parameter set, we computed the log-likelihood of the probability to make the action performed by the animal as follows:

$$L = \sum_t ln\, P_{chosen}(t) \quad (4)$$

Neurophysiological Data Analysis

Preprocessing of LFP data

LFP signals were preprocessed using a notch filter around 50Hz and a band pass filter between 1Hz and 140 Hz were applied. Artifact rejection was performed by visual inspection on the blocks of trials, keeping the ones that were not affected by the spiking activity.

Finally, filtered LFP signals were downsampled to 1000 Hz and cutted into epochs aligned on single events, namely the outcome presentation, used to define the period of analysis (from 0.0 to 0.8 sec), and the cue onset, used to define the baseline period (from -0.55 to -0.05 sec). After epoching, a second visual inspection was performed to remove artefacts from analysis, e.g. deriving from electrical interferences or by spiking activity. The period of analysis was chosen according to the fact that in some block of trials an artifact was produced at the moments of the release and of the contact between the monkey and the metal bar. Recording blocks included were cut at least 25 trials for two reasons: to be sure that the monkey discovered the correct target, and because, especially in the difficult variation of the task, we observed a decrease in performances in very late trials.

Statistical analysis of model-free and model-based behavioral correlates

In order to explore the relation between LFP power modulations and behavioral or model parameters, we computed the Linear regression between the neurophysiological signal and five behavioral variables that we considered significant for the purpose of this specific study: the RPE, the absolute value of the RPE (absRPE), the reaction time (RT), the movement time (MT) and the chosen action (Action).

As a control analysis, we assessed the degree of correlation between model-free (such as RTs and MTs) and model based (such as RPE and absRPE) behavioral variables, and single-trial LFP power values. To do so, we used a multiple linear regression (MLR) model, considering as the dependent variable (y) the average of the beta power in each trial, and as the independent variable (x) the four behavioral variables. A MLR was applied to each recording block, and once we collected all the angular coefficients relative to each block we performed a group level analysis computing a two tailed t-test. As we can see in **Table 1**, the distribution of the beta values relative to the RPE are significant for both the monkeys. This means that RPE is able to explain the variation in the beta-frequency band, and thus we focused on it to perform further analysis.

Monkey F

Regressor	t-value	p-value
RPE	5,23056	0
absRPE	3,03295	0,00276
RT	-1,88546	0,06089
MT	-0,19371	0,84661
Action	1,80368	0,07286

Monkey T

Regressor	t-value	p-value
RPE	2,61608	0,00987
absRPE	1,78841	0,07587
RT	1,47884	0,14143
MT	-3,75877	0,00025
Action	-0,34995	0,7269

Table 1: *p-values associated to the two-tailed t-test analysis of the angular coefficients resulting from the MLR*

Spectral analysis of LFP data

Time-frequency analysis

To estimate the power of the LFP signals, we performed a time-frequency analysis using the Morlet wavelet method (Cohen, 1995), considering the frequency bands from alpha to gamma - high gamma: the analysis was performed on 55 frequency steps, logarithmically spaced, in the range of 8Hz to 50Hz, and the number of cycles used for each band corresponded to its frequency divided by 4. We computed the time-frequency map in the defined periods aligned on the two previously mentioned events (baseline and period of interest). Then we applied a baseline correction at this stage of analysis, we computed the *relative change* with respect to the baseline, that corresponds to subtracting and then dividing the signal by the average over time of the baseline.

Once we obtained the corrected single-trials time-frequency maps, we divided them into two sub-datasets between rewarded and unrewarded trials. Then, we contrasted the two conditions for each monkey, performing a two-sided t-test analysis across all the trials. Thus, for each monkey we obtained a 2D p-values map with the same size of the original time-frequency maps (**Figure 2**). Since the t-test was performed

across all the considered frequencies and time points, we Bonferroni corrected the resulting p-values multiplying them by the total number of considered frequencies and time points, in order to consider the multiple comparison problem and avoid to have significative p-values by chance. The goal of this analysis was to define in which frequency band there was a peak of significant difference between the two conditions for each of the two monkeys. For monkey F, the major difference was around 25 Hz whereas for monkey T it was found at 30 Hz. These values were subsequently used to define the two frequency bands as the central frequency to estimate a single band power using the multitaper method for each monkey. Also in this case we used the relative change with respect to the baseline to correct the data.

Beta-band analyses

We then focused on a limited frequency band to study the role of beta band oscillations using the subjects' specific high beta - low gamma band power, to then perform mutual information based statistical analysis.

Single band spectral density estimation was performed using a multitaper method based on discrete prolate spheroidal (slepian) sequences (Percival and Walden, 1993; Mitra and Pesaran, 1999). To extract beta-band power estimates, LFPs time series were multiplied by k orthogonal tapers (k=4) (0.33 s in duration and 15 Hz of frequency resolution), centered at 25 and 30 Hz for monkey F and monkey T, respectively, and then Fourier-transformed.

All data analysis was performed by using subroutines written in Python (version 3.6). Data were readed and analysed using the NEO (version 0.8) and MNE (version 0.21) libraries.

Information theoretical and statistical analysis of LFP data

We used information-theoretic metrics to quantify the statistical dependency between the beta band signals and RPE signals. To this end, we computed the mutual

information (MI) between the LFP power and the behavioral variable. As a reminder, mutual information is defined as:

$$I(X;Y) = H(X) - H(X|Y)$$

Where the variables X and Yrepresent the trial by trial power of the LFP and RPEs, respectively. $H(X)$ is the entropy of X, and $H(X|Y)$is the conditional entropy of X given Y. The MI can be difficult to estimate in practice as it requires sampling the full joint distribution of the two considered variables. Therefore, here we used the recently proposed semi-parametric binning-free Gaussian-Copula Mutual Information (GCMI) (Ince et al., 2017). In short, the GCMI exploits the fact that the MI does not depend on the marginal distributions of the variables, but only on the copula function which encapsulates their statistical dependency. The GCMI is a robust rank-based approach allowing to detect any type of relation as long as this relation is roughly monotone.

For the statistical inferences, we used a group-level approach based on non-parametric permutations and encompassing non-negative measures of information (Combrisson et al., 2021) implemented in the Frites[1] Python software. To this end, we used a fixed-effect model across sessions per monkey (respectively 192 and 136 blocks for monkey F and T). By estimating the effect size across sessions, we improved the statistical power and the overall signal-to-noise ratio at the cost of ignoring the session-to-session random variations. The MI is estimated across sessions between the LFP power and the behavioral variable, at each time point and for each electrode. Finally, we used the cluster-based statistics for correcting the p-values for multiple comparisons across all time points and electrodes.

Anatomo-functional analysis of striatal territories

Once we found a strong and significant relation between beta band and RPE, we wanted to investigate how the information about the evolution of the RPE is encoded in the striatum, if its localization is restrained to the ventral striatum (Abler et al., 2006; Morris et al., 2012; Calderon et al., 2021) or if it is detectable also in dorsal

[1] https://github.com/brainets/frites

and caudal striatal regions, as was shown in previous works (Rektor et al., 2005; Valentin and O'Doherty, 2009; Asaad and Eskandar, 2011).

The electrophysiological data were collected in all the three putative regions of the striatum (i.e.,limbic, associative, and motor striatum). In order to have a better spatial resolution, we decided to divide recording sites in six different clusters following the given anatomical subdivision. The number of clusters was set according to a compromise between trial number and number of clusters.Clusters were computed using the KMeans algorithm implemented in scikit-learn, on the 3D coordinates of the electrodes defined as the antero-posterior (AP) and dorso-medial (DM) distance from the zero of the recording chamber on the monkeys' skull surface, and the depth of the electrodes. This clustering algorithm divides the data in a pre-defined number of n groups with the assumption that they should have the same variance, this result is achieved by the minimization of the within-cluster sum of squares. We also tried other techniques for clustering, but using the KMeans clustering resulted to be the best method to obtain well spatially defined and unbiased clusters.

Thus, we obtained eighteen spatial clusters and we repeated a MI based analysis similar to the one described in the previous paragraph, with the difference that a permuted MI matrix was computed for each region, and that the cluster forming threshold and the maxstat correction were applied among all this matrices. In **Figure 4,** we plotted the clusters centers' positions relative to the AP position (x axis) and the depth (y axis). In this figure, the clusters' centers are numbered following the ascending values of the average of the MI computed for each cluster, splitted up following the territory division (represented by the colours) that is used in **Figure 4**. Indeed, in **Figure 4** we can observe that higher values of MI belong to more ventromedial striatal territories, and that also the most significant values are linked to the spatial position.

After this step, we set as reference the position of the most upper and posterior recording sites and the AP and depth coordinates of each cluster center position to calculate the euclidean distance between them. This quantity gives us a good relative measure for each monkey to estimate a gradient axis. We performed a linear regression analysis between this distance and the average MI of each cluster to find out a positive correlation, suggesting that the more rostro-ventral part of the striatum

carries more information about the RPE, and that this information is not completely lost, but fading toward the caudo-dorsal part of the striatum.

Discussion

Two main aspects of the functional organization of the striatum emerge from the present study: (1) changes in LFP beta-band oscillations that may be consistent with RPE encoding (i.e., the difference between expected and actual outcomes) are observed in different parts of the striatum which are assumed to correspond to functionally distinct regions; (2) the quantity of RPE associated information is dependent on the striatal region following rostro-caudal and dorso-ventral gradients, with a maximum in the ventral part of the anterior striatum traditionally regarded as the limbic striatum in the primate. These data highlight a relationship of beta oscillatory activity in the striatum to non-motor aspects of behavior, such as the signaling of reward information, and distinct contributions for striatal regions in the evaluation of action outcome based on reward feedback.

Role of striatal beta oscillations in outcome evaluation

A key finding in our study is the occurence of LFP beta oscillations during the outcome period of the task that may play a role in evaluative processing after action choice (i.e., presence or absence of reward). Our analysis suggests that RPE was the most important variable influencing striatal LFP beta oscillations, this trend being present in data from every striatal region in which we recorded . To our knowledge, this is the first report to demonstrate that beta oscillations in the monkey striatum may play a role in RPE encoding.

Beta band oscillations in the basal ganglia are mostly associated with motor control. Indeed, numerous studies in humans and animals have provided evidence that an increased beta oscillatory activity within basal ganglia circuitry occurred with an impaired dopaminergic transmission and the expression of motor deficits observed in humans with Parkinson's disease (Brown, 2007; Jenkinson and Brown, 2011). Moreover, deep brain stimulation of the STN in dopamine-depleted conditions interferes with this abnormal oscillatory activity and improves motor symptoms (Kühn et al., 2004; Holt et al., 2019).

Beta oscillations have also been reported in the striatal LFP activity of normal animals, both rodents and monkeys, during specific phases of behavioral tasks (Berke et al., 2004; Courtemanche et al., 2003; Leventhal et al., 2012; Schmidt et al.,

2013; Bartolo et al., 2014), but the potential functional significance of such oscillatory activities is still under debate. In particular, despite the proposed role of the striatum in action valuation and reward-driven learning, few studies have specifically investigated whether striatal beta oscillations could possibly be associated with reward processing (Howe et al., 2011; Leventhal et al., 2012; Münte et al., 2017; Schwerdt et al., 2020). For example, the work of Leventhal et al. (2012) has shown that beta band oscillations are associated with cue utilization in rat striatum. They used four different variants of the classic Go-NoGo task, founding a whole-striatum and non lateralized event-related synchronization (ERS) in the beta band associated to the cue, and not linked to motor initiation or suppression, in every variant of this task. The relevant feature that should follow the cue to produce a beta ERS is the presence of the reward. Indeed in all of these task variants, in which the reward is deterministic, if we think about the cognitive role of the cue producing the beta band power increase, it seems 'anticipating' the reward release.

Reward prediction error encoding in the striatum

The role of midbrain dopamine neurons in RPE encoding is well established (Fiorillo et al., 2003; Abler et al., 2006; Bray and O'Doherty, 2007; Fujiyama et al., 2015). Animal electrophysiology and human neuroimaging have provided extensive evidence of RPE-related activity in the striatum (Apicella et al., 2009; Roesch et al., 2009; Oyama et al., 2010; Asaad and Eskandar, 2011; Stalnaker et al., 2012) which is the main target structure of ascending dopamine projections from neurons located in the substantia nigra *pars compacta* and ventral tegmental area. RPE is a non-linear measure that can have positive or negative values, computed as the value of the reward (0 or 1) minus the value of the prediction relative to the state-action couple (Asaad and Eskandar, 2011); it can't be directly measured, for this reason we used a Q-learning model fitted on monkeys' choice behavior to compute it trial by trial. RPE is essential for adaptive behavior in order to avoid non rewarding actions and exploit the rewarding ones, by improving the predictions about future outcomes (O'Doherty et al., 2017), playing a crucial role in the acquisition of new learned behaviors (Ressler, 2004; O'Doherty, 2007; Keramati et al., 2011; Nonomura et al., 2018). From our work, a significant increment of mutual information between the beta band and the RPE is detected in both monkeys, with a slightly stronger effect in

monkey F compared to monkey T. To interpret this result, we should consider that the MI between two variables can be considered as an index of covariation between them. Thus, in this analysis an increment in MI corresponds to a strong covariation between the across trial evolution of the beta-oscillations power and the RPE. Moreover, according to the statistic we used, the significance indicates that these variables covariates between them over a substantial number or recording blocks.
Thus, the striatum can have a major role in encoding and transmission of RPE signals across different functional regions. More studies about the transmission of RPE signals both intra-striatum and across the striato-cortical network are needed in order to better understand the time course, the localization and the behavioral salience of this signal, so important for the regulation of higher cognitive processes.
Finally, one may consider that the observed changes in striatal beta activity could possibly be associated, at least in part, with other aspects of information processing during the outcome period of the choice task, such as return movements to the resting bar or the experience during reward consumption (sensory pleasure or mouth movements). Additional studies are necessary to disambiguate the affective, motor, or cognitive origin of changes in beta oscillations at the end of the trial in our task.

Functional parcellation of the striatum

Different parts of the striatum and their corresponding cortical inputs are assumed to serve different functions, with a major involvement of the dorsal part of the posterior putamen in motor processing, whereas the ventral part of the anterior caudate nucleus and putamen is more concerned with mediating motivation and reward (Apicella et al., 1991; Fiorillo et al., 2003; Marchand et al., 2008; Brovelli et al., 2011; Pennartz et al., 2011; Schultz, 2016a, 2016b; Han et al., 2021). Taking account of these regional differences, we investigated LFP activity in both anterior and posterior parts of the striatum. According to our results, different clusters of recording sites were associated with different quantities of MI in the outcome period. Thus we wanted to understand if the total value of MI summed over time is following a spatial organization, to then perform a linear regression analysis between the total MI and their relative position to a rostro-ventral to caudo-dorsal anatomical axis. We chose to form the clusters respecting the classical functional striatal regionalization given by well known anatomical constraints (Jahanshahi et al., 2015).

Several lines of evidence point to a major involvement of the ventral part of the anterior striatum, including the nucleus accumbens, in the processing of reward-related information (Apicella et al., 1991; O'Doherty, 2004; Schultz, 2016c). In particular, a number of studies have highlighted the role of the ventral striatum in the computation of RPEs (Abler et al., 2006; Bray and O'Doherty, 2007; Schultz, 2016a; Calderon et al., 2021). Our results suggest that the information about RPE is present, to varying degrees, in all parts of the explored striatum, forming a fading gradient stronger toward the rostro-ventral striatum and weaker toward its caudo-dorsal part. This result is in line with fMRI studies in humans showing that striatal circuitry is able to establish different functional gradients, spanning from the dopaminergic signaling to the cognitive control (Mestres-Missé et al., 2012; Vogelsang and D'Esposito, 2018; Alberquilla et al., 2020; Han et al., 2021), determined by parallel pathways from motor, associative, and limbic cortical areas running through different regions of the striatum.

It seems that a communication between distinct functional territories is implemented by gradients of information carried by oscillations. Indeed, also if a well structured connectivity is needed to transmit precise signals, the information contained in those can participate in other behavioral functions. RPE is indeed needed to update the inner model of action values in response to a particular state, and those values should be retained in short term memory in order to plan future actions in a goal-directed way. Given its intricate internal connectivity shaped by cholinergic and GABAergic interneurons, and its diffuse projections over cortical and subcortical regions, the striatum lends itself well to the role of messenger. Thus, the RPE gradient can be a result of the internal striatal transmission (and processing) of the dopaminergic signal, allowing it to reach different behavioral systems. Our results are in line with the idea that the RPE is an important signal affecting several aspects of the behavior, and that for this reason it should propagate in limbic, cognitive and motor areas of the brain (Silvetti et al., 2014; Schultz, 2016b). .

We have already pointed out that the RPE signal exerts a driving influence on goal-directed learning. As such, it is used together with our present knowledge in order to plan future actions (Takikawa et al., 2002; Ressler, 2004; Gläscher et al., 2010; Izawa and Shadmehr, 2011; Schultz, 2016a). Thus, one can expect that this

signal should be able to reach all striatal regions in order to participate in limbic, associative and motor functions, and propagate in the functional associated cortices, such as the cingulate cortex, the prefrontal cortex and the premotor cortex (Oya et al., 2005; Mestres-Missé et al., 2012; Vogelsang and D'Esposito, 2018).

In the present study, we focused on LFPs oscillations to study their implication in outcome processing. Contrary to spiking activity that is detected at higher frequencies reflecting the very local activation of neurons, LFPs are detected at lower frequencies, and are assumed to reflect the activity of populations of neurons (Buzsáki et al., 2012). Thus, LFPs can be considered as signals recorded on a relatively larger area (generally a couple of millimeters of diameters from the electrode), containing the average coordinate activity of several neurons. According to literature, the main contributors to LFPs are the excitatory and inhibitory postsynaptic potentials (E/IPSP), and sometimes also membrane hyperpolarization (Buzsáki, 2006; van der Meer, 2010; Buzsáki et al., 2012). The recorded activity can contain rhythmic oscillation in specific frequency bands, which can be related to some environmental, behavioral or cognitive aspects. Although this signal can sometimes contain traces of leaking activity from surrounding brain areas, LFPs are increasingly used for the study of striatal activity (Courtemanche et al., 2003; Berke et al., 2004; Brown and Williams, 2005; van der Meer, 2009; Münte et al., 2017; Suzuki and Tanaka, 2019). Taking into account this limitation, one of our future interests will be to consider the Spiking-LFP coupling to better investigate the role of beta oscillations in encoding RPE and its distribution through striatum. This study could be also helpful to understand differences and similarities between rodents' and primates' striatal activity.

Indeed, although we have concentrated, in our study, on striatal LFP oscillations in the beta band, experiments with rodents have shown that LFP oscillations in the gamma band are more prominent in the ventromedial striatum, as compared to the dorsolateral striatum (Berke et al. 2004; Berke 2009; van der Meer and Redish 2009; van der Meer et al. 2010; Kalensher et al. 2010). To our knowledge, there is no evidence of a similar RPE related gamma activity in primates striatum. Given their similar role in encoding reward related information and outlining a gradient of activity, we wonder if this difference in bands' activity can be given by an interspecific shift in oscillations, consequent to morphological striatal change.

Conclusion

The aim of this study was to assess the role of different functional striatal regions in encoding RPEs signals. To do so, we analysed LFPs data recorded in three different striatal anatomical regions of two monkeys while performing a free choice probabilistic learning task. We provided new evidence that changes in beta band oscillations may reflect the encoding of RPEs defined in reinforcement learning models. Then, we divided the recording sites in eighteen spatial clusters and we observed that such changes were dominant in the rostro-ventral rather than the caudo-dorsal striatum, supporting the notion of a prominent role for the limbic part of the striatum in evaluative processing useful for future actions. Based on our mapping of the spatial organization of oscillatory beta activity in the striatum, we propose that the RPE encoding can occur first in the ventral region and then spread in the dorsal region. This finding may be of clinical importance as it is known that dorsal and ventral parts of the striatum are differentially involved in neuropsychiatric diseases, with dorsal striatal circuits mainly related to motor and cognitive disorders, whereas ventral striatal circuits are involved rather in the expression of affective disorders and compulsive behaviors. However, more studies are needed to understand which are the neural computations at the base of striatal gradients formation.

Bibliography

- Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. NeuroImage *31*, 790–795.
- Alberquilla, S., Gonzalez-Granillo, A., Martín, E.D., and Moratalla, R. (2020). Dopamine regulates spine density in striatal projection neurons in a concentration-dependent manner. Neurobiol. Dis. *134*, 104666.
- Apicella, P., Ljungberg, T., Scarnati, E., and Schultz, W. (1991). Responses to reward in monkey dorsal and ventral striatum. Exp. Brain Res. *85*.
- Apicella, P., Deffains, M., Ravel, S., and Legallet, E. (2009). Tonically active neurons in the striatum differentiate between delivery and omission of expected reward in a probabilistic task context. Eur. J. Neurosci. *30*, 515–526.
- Asaad, W.F., and Eskandar, E.N. (2011). Encoding of Both Positive and Negative Reward Prediction Errors by Neurons of the Primate Lateral Prefrontal Cortex and Caudate Nucleus. J. Neurosci. *31*, 17772–17787.
- Balleine, B.W., Delgado, M.R., and Hikosaka, O. (2007). The Role of the Dorsal Striatum in Reward and Decision-Making. J. Neurosci. *27*, 8161–8165.
- Banaie Boroujeni, K., Oemisch, M., Hassani, S.A., and Womelsdorf, T. (2020). Fast spiking interneuron activity in primate striatum tracks learning of attention cues. Proc. Natl. Acad. Sci. *117*, 18049–18058.
- Bartolo, R., Prado, L., and Merchant, H. (2014). Information Processing in the Primate Basal Ganglia during Sensory-Guided and Internally Driven Rhythmic Tapping. J. Neurosci. *34*, 3910–3923.
- Berke, J.D., Okatan, M., Skurski, J., and Eichenbaum, H.B. (2004). Oscillatory Entrainment of Striatal Neurons in Freely Moving Rats. 14.
- Bray, S., and O'Doherty, J. (2007). Neural Coding of Reward-Prediction Error Signals During Classical Conditioning With Attractive Faces. J. Neurophysiol. *97*, 3036–3045.
- Brovelli, A., Nazarian, B., Meunier, M., and Boussaoud, D. (2011). Differential roles of caudate nucleus and putamen during instrumental learning. NeuroImage *57*, 1580–1590.
- Brown, P. (2007). Abnormal oscillatory synchronisation in the motor system leads to impaired movement. Curr. Opin. Neurobiol. *17*, 656–664.
- Brown, P., and Williams, D. (2005). Basal ganglia local field potential activity: Character and functional significance in the human. Clin. Neurophysiol. *116*, 2510–2519.
- Buzsáki, G. (2006). Rhythms of the Brain (Oxford University Press).
- Buzsáki, G., Anastassiou, C.A., and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. *13*, 407–420.
- Calderon, C.B., De Loof, E., Ergo, K., Snoeck, A., Boehler, C.N., and Verguts, T. (2021). Signed Reward Prediction Errors in the Ventral Striatum Drive Episodic Memory. J. Neurosci. *41*, 1716–1726.
- Cohen, L. (1995). Time-frequency analysis (Englewood Cliffs, N.J: Prentice Hall PTR).
- Combrisson, E., Allegra, M., Basanisi, R., Ince, R.A.A., Giordano, B., Bastin, J., and Brovelli, A. (2021). Group-level inference of information-based

measures for the analyses of cognitive brain networks from neurophysiological data (Neuroscience).
- Courtemanche, R., Fujii, N., and Graybiel, A.M. (2003). Synchronous, Focally Modulated ℕ-Band Oscillations Characterize Local Field Potential Activity in the Striatum of Awake Behaving Monkeys. 12.
- Deffains, M., Iskhakova, L., Katabi, S., Haber, S.N., Israel, Z., and Bergman, H. (2016). Subthalamic, not striatal, activity correlates with basal ganglia downstream activity in normal and parkinsonian monkeys. ELife *5*.
- Delgado, M.R., Miller, M.M., Inati, S., and Phelps, E.A. (2005). An fMRI study of reward-related probability learning. NeuroImage *24*, 862–873.
- Eisinger, R.S., Urdaneta, M.E., Foote, K.D., Okun, M.S., and Gunduz, A. (2018). Non-motor Characterization of the Basal Ganglia: Evidence From Human and Non-human Primate Electrophysiology. Front. Neurosci. *12*.
- Feingold, J., Gibson, D.J., DePasquale, B., and Graybiel, A.M. (2015). Bursts of beta oscillation differentiate postperformance activity in the striatum and motor cortex of monkeys performing movement tasks. Proc. Natl. Acad. Sci. *112*, 13687–13692.
- Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. Science *299*, 1898–1902.
- Fujiyama, F., Takahashi, S., and Karube, F. (2015). Morphological elucidation of basal ganglia circuits contributing reward prediction. Front. Neurosci. *9*.
- Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron *66*, 585–595.
- Goldberg, J.A. (2004). Spike Synchronization in the Cortex-Basal Ganglia Networks of Parkinsonian Primates Reflects Global Dynamics of the Local Field Potentials. J. Neurosci. *24*, 6003–6010.
- Han, M.-J., Park, C.-U., Kang, S., Kim, B., Nikolaidis, A., Milham, M.P., Hong, S.J., Kim, S.-G., and Baeg, E. (2021). Mapping functional gradients of the striatal circuit using simultaneous microelectric stimulation and ultrahigh-field fMRI in non-human primates. NeuroImage *236*, 118077.
- Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., and Rangel, A. (2008). Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. J. Neurosci. *28*, 5623–5630.
- Holt, A.B., Kormann, E., Gulberti, A., Pötter-Nerger, M., McNamara, C.G., Cagnan, H., Baaske, M.K., Little, S., Köppen, J.A., Buhmann, C., et al. (2019). Phase-Dependent Suppression of Beta Oscillations in Parkinson's Disease Patients. J. Neurosci. *39*, 1119–1134.
- Howe, M.W., Atallah, H.E., McCool, A., Gibson, D.J., and Graybiel, A.M. (2011). Habit learning is associated with major shifts in frequencies of oscillatory activity and synchronized spike firing in striatum. Proc. Natl. Acad. Sci. *108*, 16801–16806.
- Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., and Schyns, P.G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula: Gaussian Copula Mutual Information. Hum. Brain Mapp. *38*, 1541–1573.
- Ito, M., and Doya, K. (2009). Validation of Decision-Making Models and Analysis of Decision Variables in the Rat Basal Ganglia. J. Neurosci. *29*, 9861–9874.

- Izawa, J., and Shadmehr, R. (2011). Learning from Sensory and Reward Prediction Errors during Motor Adaptation. PLoS Comput. Biol. *7*, e1002012.
- Jahanshahi, M., Obeso, I., Rothwell, J.C., and Obeso, J.A. (2015). A fronto–striato–subthalamic–pallidal network for goal-directed and habitual inhibition. Nat. Rev. Neurosci. *16*, 719–732.
- Jenkinson, N., and Brown, P. (2011). New insights into the relationship between dopamine, beta oscillations and motor function. Trends Neurosci. *34*, 611–618.
- Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. PLoS Comput. Biol. *7*, e1002055.
- Kondabolu, K., Roberts, E.A., Bucklin, M., McCarthy, M.M., Kopell, N., and Han, X. (2016). Striatal cholinergic interneurons generate beta and gamma oscillations in the corticostriatal circuit and produce motor deficits. Proc. Natl. Acad. Sci. *113*, E3159–E3168.
- Kühn, A.A., Williams, D., Kupsch, A., Limousin, P., Hariz, M., Schneider, G., Yarrow, K., and Brown, P. (2004). Event‐related beta desynchronization in human subthalamic nucleus correlates with motor performance. Brain *127*, 735–746.
- Kumar, P., Goer, F., Murray, L., Dillon, D.G., Beltzer, M.L., Cohen, A.L., Brooks, N.H., and Pizzagalli, D.A. (2018). Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. Neuropsychopharmacology *43*, 1581–1588.
- Lanciego, J.L., Luquin, N., and Obeso, J.A. (2012). Functional Neuroanatomy of the Basal Ganglia. Cold Spring Harb. Perspect. Med. *2*, a009621–a009621.
- Lau, B., and Glimcher, P.W. (2007). Action and Outcome Encoding in the Primate Caudate Nucleus. J. Neurosci. *27*, 14502–14514.
- Leventhal, D.K., Gage, G.J., Schmidt, R., Pettibone, J.R., Case, A.C., and Berke, J.D. (2012). Basal Ganglia Beta Oscillations Accompany Cue Utilization. Neuron *73*, 523–536.
- Liljeholm, M., and O'Doherty, J.P. (2012). Contributions of the striatum to learning, motivation, and performance: an associative account. Trends Cogn. Sci. *16*, 467–475.
- Marchand, W.R., Lee, J.N., Thatcher, J.W., Hsu, E.W., Rashkin, E., Suchy, Y., Chelune, G., Starr, J., and Barbera, S.S. (2008). Putamen coactivation during motor task execution. NeuroReport *19*, 957–960.
- Marche, K., and Apicella, P. (2021). Activity of fast‐spiking interneurons in the monkey striatum during reaching movements guided by external cues or by a free choice. Eur. J. Neurosci. *53*, 1752–1768.
- Marche, K., Martel, A.-C., and Apicella, P. (2017). Differences between Dorsal and Ventral Striatum in the Sensitivity of Tonically Active Neurons to Rewarding Events. Front. Syst. Neurosci. *11*.
- van der Meer, M. (2010). Integrating early results on ventral striatal gamma oscillations in the rat. Front. Neurosci.
- van der Meer, M.A.A. (2009). Low and high gamma oscillations in rat ventral striatum have distinct relationships to behavior, reward, and spiking activity on a learned spatial decision task. Front. Integr. Neurosci. *3*.
- Mestres-Missé, A., Turner, R., and Friederici, A.D. (2012). An anterior–posterior gradient of cognitive control within the dorsomedial striatum. NeuroImage *62*, 41–47.

- Mitra, P.P., and Pesaran, B. (1999). Analysis of Dynamic Brain Imaging Data. Biophys. J. *76*, 691–708.
- Morris, R.W., Vercammen, A., Lenroot, R., Moore, L., Langton, J.M., Short, B., Kulkarni, J., Curtis, J., O'Donnell, M., Weickert, C.S., et al. (2012). Disambiguating ventral striatum fMRI-related bold signal during reward prediction in schizophrenia. Mol. Psychiatry *17*, 280–289.
- Münte, T.F., Marco-Pallares, J., Bolat, S., Heldmann, M., Lütjens, G., Nager, W., Müller-Vahl, K., and Krauss, J.K. (2017). The human globus pallidus internus is sensitive to rewards – Evidence from intracerebral recordings. Brain Stimulat. *10*, 657–663.
- Nakano, K., Kayahara, T., Tsutsumi, T., and Ushiro, H. (2000). Neural circuits and functional organization of the striatum. J. Neurol. *247*, V1–V15.
- Nini, A., Feingold, A., Slovin, H., and Bergman, H. (1995). Neurons in the globus pallidus do not show correlated activity in the normal monkey, but phase-locked oscillations appear in the MPTP model of parkinsonism. J. Neurophysiol. *74*, 1800–1805.
- Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., Watanabe, M., Yamanaka, K., Enomoto, K., Chiken, S., et al. (2018). Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. Neuron *99*, 1302-1314.e5.
- O'Doherty, J.P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr. Opin. Neurobiol. *14*, 769–776.
- O'Doherty, J.P. (2007). Lights, Camembert, Action! The Role of Human Orbitofrontal Cortex in Encoding Stimuli, Rewards, and Choices. Ann. N. Y. Acad. Sci. *1121*, 254–272.
- O'Doherty, J.P., Hampton, A., and Kim, H. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. Ann. N. Y. Acad. Sci. *1104*, 35–53.
- O'Doherty, J.P., Cockburn, J., and Pauli, W.M. (2017). Learning, reward, and decision making. Annu. Rev. Psychol. *68*, 73–100.
- Oya, H., Adolphs, R., Kawasaki, H., Bechara, A., Damasio, A., and Howard, M.A. (2005). Electrophysiological correlates of reward prediction error recorded in the human prefrontal cortex. Proc. Natl. Acad. Sci. *102*, 8351–8356.
- Oyama, K., Hernadi, I., Iijima, T., and Tsutsui, K.-I. (2010). Reward Prediction Error Coding in Dorsal Striatal Neurons. J. Neurosci. *30*, 11447–11457.
- Parent, A. (1990). Extrinsic connedions of the basal ganglia. Trends Neurosci. *13*, 254–258.
- Park, S.Q., Kahnt, T., Talmi, D., Rieskamp, J., Dolan, R.J., and Heekeren, H.R. (2012). Adaptive coding of reward prediction errors is gated by striatal coupling. Proc. Natl. Acad. Sci. *109*, 4285–4289.
- Pennartz, C.M.A., Ito, R., Verschure, P.F.M.J., Battaglia, F.P., and Robbins, T.W. (2011). The hippocampal–striatal axis in learning, prediction and goal-directed behavior. Trends Neurosci. *34*, 548–559.
- Percival, D.B., and Walden, A.T. (1993). Spectral Analysis for Physical Applications (Cambridge University Press).
- Rektor, I., Bareš, M., Brázdil, M., Kaňovský, P., Rektorová, I., Soch??rková, D., Kubová, D., Kuba, R., and Daniel, P. (2005). Cognitive- and

movement-related potentials recorded in the human basal ganglia. Mov. Disord. *20*, 562–568.
- Ressler, N. (2004). Rewards and punishments, goal-directed behavior and consciousness. Neurosci. Biobehav. Rev. *28*, 27–39.
- Roesch, M.R., Singh, T., Brown, P.L., Mullins, S.E., and Schoenbaum, G. (2009). Ventral Striatal Neurons Encode the Value of the Chosen Action in Rats Deciding between Differently Delayed or Sized Rewards. J. Neurosci. *29*, 13365–13376.
- Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. Science *310*, 1337–1340.
- Schmidt, R., Leventhal, D.K., Mallet, N., Chen, F., and Berke, J.D. (2013). Canceling actions involves a race between basal ganglia pathways. Nat. Neurosci. *16*, 1118–1124.
- Schultz, W. (2007). Multiple Dopamine Functions at Different Time Courses. Annu. Rev. Neurosci. *30*, 259–288.
- Schultz, W. (2016a). Dopamine reward prediction-error signalling: a two-component response. Nat. Rev. Neurosci. *17*, 183–195.
- Schultz, W. (2016b). Dopamine reward prediction error coding. Dialogues Clin. Neurosci. *18*, 10.
- Schultz, W. (2016c). Reward functions of the basal ganglia. J. Neural Transm. *123*, 679–693.
- Schwerdt, H.N., Amemori, K., Gibson, D.J., Stanwicks, L.L., Yoshida, T., Bichot, N.P., Amemori, S., Desimone, R., Langer, R., Cima, M.J., et al. (2020). Dopamine and beta-band oscillations differentially link to striatal value and motor control. Sci. Adv. *6*, eabb9226.
- Seo, M., Lee, E., and Averbeck, B.B. (2012). Action Selection and Action Value in Frontal-Striatal Circuits. Neuron *74*, 947–960.
- Silvetti, M., Nuñez Castellar, E., Roger, C., and Verguts, T. (2014). Reward expectation and prediction error in human medial frontal cortex: An EEG study. NeuroImage *84*, 376–382.
- Stalnaker, T.A., Calhoon, G.G., Ogawa, M., Roesch, M.R., and Schoenbaum, G. (2012). Reward Prediction Error Signaling in Posterior Dorsomedial Striatum Is Action Specific. J. Neurosci. *32*, 10296–10305.
- Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: an introduction.
- Suzuki, T.W., and Tanaka, M. (2019). Neural oscillations in the primate caudate nucleus correlate with different preparatory states for temporal production. Commun. Biol. *2*.
- Takikawa, Y., Kawagoe, R., Itoh, H., Nakahara, H., and Hikosaka, O. (2002). Modulation of saccadic eye movements by predicted reward outcome. Exp. Brain Res. *142*, 284–291.
- Valentin, V.V., and O'Doherty, J.P. (2009). Overlapping Prediction Errors in Dorsal Striatum During Instrumental Learning With Juice and Money Reward in the Human Brain. J. Neurophysiol. *102*, 3384–3391.
- Vogelsang, D.A., and D'Esposito, M. (2018). Is There Evidence for a Rostral-Caudal Gradient in Fronto-Striatal Loops and What Role Does Dopamine Play? Front. Neurosci. *12*.
- Wang, K.S., Smith, D.V., and Delgado, M.R. (2016). Using fMRI to study reward processing in humans: past, present, and future. J. Neurophysiol. *115*, 1664–1678.
- Watkins, C.J., and Dayan, P. (1992). Q-learning. Mach. Learn. *8*, 279–292.

- Wichmann, T., Bergman, H., and DeLong, M.R. (1994). The primate subthalamic nucleus. I. Functional properties in intact animals. J. Neurophysiol. *72*, 494–506.
- Yamada, H., Inokawa, H., Matsumoto, N., Ueda, Y., Enomoto, K., and Kimura, M. (2013). Coding of the long-term value of multiple future rewards in the primate striatum. J. Neurophysiol. *109*, 1140–1151.

Section 4. Dynamics of human cortical circuits mediating goal-directed causal learning

Dynamics of human cortical circuits mediating goal-directed causal learning

Basanisi R.[1], Combrisson E.[1], Dauce E.[1], Brovelli A.[1]

1 Institut de Neurosciences de la Timone, Aix Marseille Université, UMR 7289 CNRS, 13005, Marseille, France

Corresponding authors:
Ruggero Basanisi
ruggero.basanisi@gmail.com
Institut de Neurosciences de la Timone (INT),
UMR 7289 CNRS, Aix Marseille University,
Campus de Santé Timone,
27 Bd. Jean Moulin,
13385 Marseille, France

Andrea Brovelli
andrea.brovelli@univ-amu.fr
Institut de Neurosciences de la Timone (INT),
UMR 7289 CNRS, Aix Marseille University,
Campus de Santé Timone,
27 Bd. Jean Moulin,
13385 Marseille, France

Keywords:
Cortical dynamics, magnetoencephalography, MEG, contingency, causal learning, goal-directed learning, learning, Bayesian model.

Abstract

Humans have an extraordinary ability to infer causal relations between actions, or more generally behaviors, and their consequences. Such sense of causality is thought to be linked to the action-outcome contingency, which is defined as the difference between the probability of observing a given outcome when an action is performed (P(O|A) and the probability of receiving the same outcome when the action is withheld (P(O|¬A). Although neural correlates of goal-directed causal learning are well addressed in literature, less is still known about the temporal dynamics of the underlying brain regions. We analysed the cortical high-gamma activity (HGA, 60-120Hz) estimated from magnetoencephalography (MEG) data recorded from human participants while performing a causal learning task with different associated contingency values. A Bayesian ideal observer model was used to estimate the evolution of action-outcome probabilities and contingencies from the sequence of stimuli and behavioral choices. Model-based analysis of HGA exploiting information theory measures combined with cluster-based statistics was used to identify the brain regions mediating such learning computations, together with their temporal dynamics. Our findings suggest a major role of the prefrontal cortices, and in particular the orbitofrontal cortex (OFC) and the rostral prefrontal cortices in encoding post outcome signals related to contingency update and learning.

Introduction

In the context of goal-directed learning (Balleine and Dickinson, 1998; Dickinson and Balleine, 2000; Dolan and Dayan, 2013), mammals and especially humans prove to be particularly able to infer causal relations between the actions they perform and the outcomes they receive (Blaisdell, 2006; Penn and Povinelli, 2007; Liljeholm, 2018, 2021). This ability is required in order to manage a probabilistic environment and to flexibly adapt to changing rules. The ability to infer causal relations relies on the ability of creating an inner representation of the environment, and to use this internal model of actions-outcomes interactions to compute their contingency value (Dickinson and Balleine, 2000; Moore et al., 2009). Psychologists defined instrumental contingency using a mathematical formulation, according to which the perceived contingency value, called ΔP, corresponds to the difference between the conditional probability of obtaining an outcome after performing an action (P(O|A)) and the conditional probability of receiving the same outcome when the action is withheld (P(O|¬A)) (Hammond, 1980; Allan and Jenkins, 1980; Allan, 1993; Allan et al., 2008; Morris et al., 2017).

Thus the contingency value corresponds to a subjective judgement of causality, that according to its definition can take values from -1 to 1, being the difference between two probability values ranging from 0 to 1. A positive ΔP corresponds to a positive causal perception meaning that the subject has the impression that the action triggers the outcome, while a negative ΔP value corresponds to a negative causal perception, meaning that the subject has the impression that the action prevents the outcome. Additionally, when the ΔP value is close to 0 the subject will have a null causal perception, meaning that there is no apparent causal relation between the action and the outcome (Shanks and Dickinson, 1991; Msetfi et al., 2013).

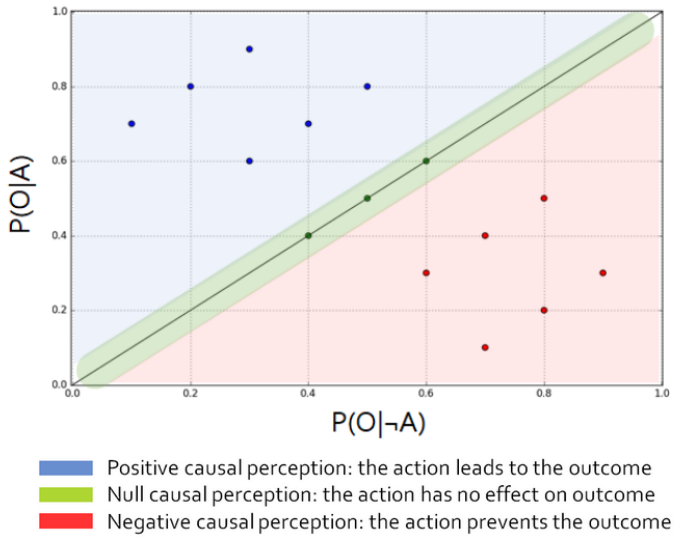
Previous behavioral studies, in which different couples of the two conditional probabilities were used, confirmed that humans are sensitive to small variations of ΔP, and that this measure reflects causal judgment (Wasserman et al., 1983; Shanks, 1985). Other studies used a contingency degradation paradigm (Balleine and Dickinson, 1998), in which the value of P(O|A) is fixed and the value of P(O|¬A) is

gradually increased leading the subject to a loss of interest toward the action, confirming sensitivity toward action-outcomes contingency values. Studies both in rats (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Yin et al., 2005) and in humans (Tanaka et al., 2008) used this paradigm to show the important role of cortical prefrontal regions such as the prelimbic cortex, the medial prefrontal cortex (mPFC) and the medial orbitofrontal cortex (mOFC), together with subcortical regions such as the dorsal striatum. Importantly, some functional Magnetic Resonance Imaging (fMRI) study (Liljeholm et al., 2011, 2013) found significant correlations between distinct aspects of contingency learning and cortical and subcortical regions, highlighting the implication of more posterior areas such as the superior and inferior parietal lobule.

Although causal learning is strictly linked to instrumental learning and goal-directed learning, of which brain areas implication are widely addressed by literature, less is still known about the link between neural and computational dynamic underlying causal learning. The aim of this study is to assess the temporal dynamics of the contribution of different cortical brain regions during contingency acquisition in humans.

To do so, we asked eighteen participants to perform a goal-directed causal learning task while their brain activity was recorded with magnetoencephalography (MEG), in order to extract and analyse the power of the signal in the high-gamma band, that is known to account for local computations (von Stein and Sarnthein, 2000; Buzsáki and Wang, 2012) and reflect fMRI hemodynamic response (Logothetis et al., 2001; Brovelli et al., 2005). Then, in order to estimate the trial-by-trial evolution of the conditional probabilities, we built a Bayesian computational model of an optimal observer (Meyniel et al., 2015). We fitted the model with the behavior of the participants in order to obtain the subjective progression of the perception of task related variables, such as the ΔP. We then used mutual information (MI) to investigate the relations between cortical activity and behavior and we analysed the results using a non-parametric cluster based statistics method.

Methods



Scenario	P(O|A)	P(O|¬A)	ΔP
1	0.7	0.4	0.3
2	0.4	0.7	-0.3
3	0.7	0.1	0.6
4	0.2	0.8	-0.6
5	0.4	0.4	0
6	0.1	0.7	-0.6
7	0.6	0.3	0.3
8	0.8	0.5	0.3
9	0.6	0.6	0
10	0.3	0.9	-0.6
11	0.9	0.3	0.6
12	0.3	0.6	-0.3
13	0.5	0.8	-0.3
14	0.8	0.2	0.6
15	0.5	0.5	0

Figure 1. *Probability values linked to the volleyball task. The causal perception ΔP is given by the difference between the contingency values P(O|A) and P(O|¬A). A positive ΔP is linked to a positive causal perception, a negative ΔP is linked to a negative causal perception, while a ΔP close to zero is linked to no causal perception between the action and the outcome.*

Experimental set-up and causal learning task

Eighteen healthy participants accepted to take part in our study, all of them were right handed, 13 were females and 5 males, and the average age was around 25 years. We submitted to them a written informed consent according to established institutional guidelines and local ethics committee. At the end of the experience, participants received a 50€ monetary compensation.

We designed an original task, that we called the 'volleyball task', which allowed us to modulate both the actions-outcomes conditional probabilities (P(O|A) and P(O|¬A)) in order to obtain 5 possible values of ΔP, both positive and negative (-0.6; -0.3; 0; 0.3; 0.6). To avoid introducing some possible biases associated with the values of conditional probabilities, each of these ΔP values was computed using three different couples of P(O|A) and P(O|¬A); thus we obtained a total of fifteen possible scenarios (3 couples × 5 ΔP). A list of the fifteen scenarios with their associated probabilities,

and the resulting contingency values can be found in **Figure 1**. Moreover, as shown in **Figure 1**, introducing this variability in the task design is also important to study different intensities of causal perception.

Participants performed fifteen scenarios of the task in a randomized order, and all of them received the same instructions. Participants were instructed to impersonate a volleyball trainer, trying to evaluate the causal effect of fifteen players in their team. To do so, they had the opportunity to simulate forty matches for each player (corresponding to forty trials for each scenario). This task was self-paced, meaning that no previous stimulus about the beginning of trials was given to participants: they could have started a trial in every moment by performing a motor response. Thus, when they wanted to simulate a match, they could have chosen if to let the questioned player play the match or not, by pushing one of two buttons under their right hand. Each button was associated to a visual cue (a 'play' or a 'pause' symbol) projected on a black screen informing the participants about their corresponding action value The order of the cues (and as a consequence the action associated to the buttons) was inverted in half of the scenarios to avoid possible biases due to positional effect. The outcome of the match was presented 250 msec after the choice of the participant. The feedback could be either a green happy face or a red sad face appeared at the center of the screen to inform the participants about the result of the match (respectively win or lose). The outcome image was displayed for 1.5 sec, during which it was not possible to perform an action. After the outcome image disappeared, we imposed an additional waiting period of 300 msec before taking any other action in consideration. See **Figure 2** for a visual description of a single trial time course. At the end of each scenario, we asked the participants to verbally report a 'causal score' from -100 to 100 to evaluate the performances of the player, where -100 corresponds to a very negative causal perception ('everytime I put this player in my team they lose'), 100 corresponds to a very positive causal perception ('everytime I put this player in my team they win') and 0 corresponds to a null causal perception ('the player doesn't affect at all the performances of my team')

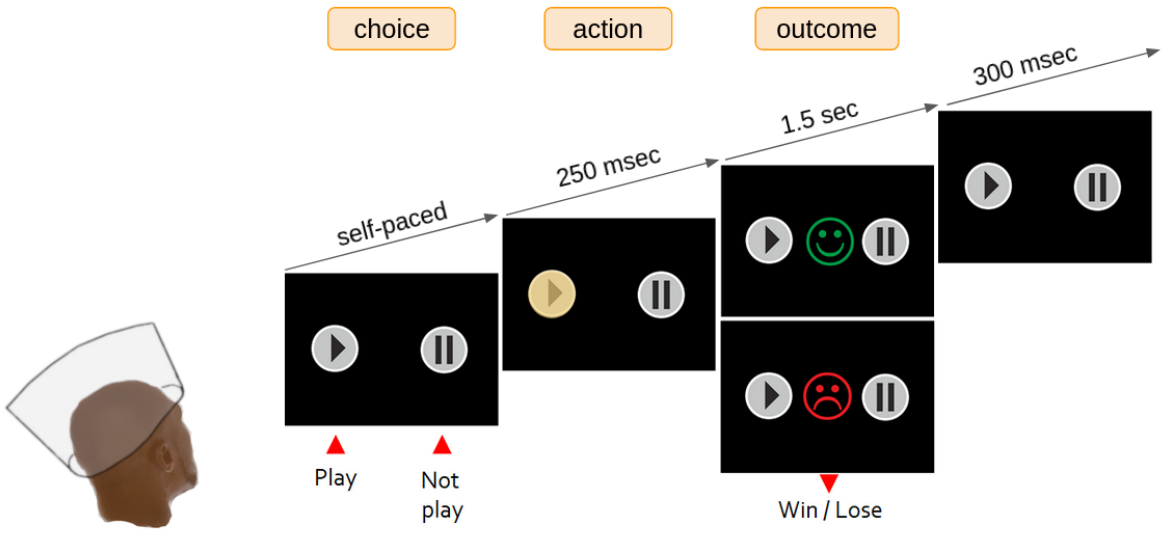


Figure 2. *Single trial description of the volleyball task. The task was self-paced, meaning that the participants didn't receive any cue to start. 250 msec after selection and execution of the action ('play' or 'not play') they received an outcome. The outcome was informative about the result of the simulated match and lasted 1.5 sec. Only after an additional 300 msec the participants were able to start the next trial.*

Data acquisition

Anatomical MRI images were acquired for each participant using a 3T whole-body imager equipped with a circular polarized head coil. MEG recordings were performed using a 248 magnetometers system (4D Neuroimaging, magnes 3600). Five additional electrodes were placed to record cardiac activity, eye-blinks and both vertical and horizontal eye movements. Visual stimuli were projected using a video projection, and motor responses were acquired using a LUMItouch optical response keypad with five keys. Presentation software was used for stimulus delivery and experimental control during MEG acquisition. Sampling rate was 2034.5 Hz. We recorded as a baseline ten seconds of resting state activity at the beginning of each scenario, asking the participants to keep their eyes open and fixate a red cross in the middle of a black screen. The fifteen scenarios were divided in five recording blocks to offer participants the opportunity to have pauses. Location of the participant's head with respect to the MEG sensors was recorded both at the beginning and end of each recording block to potentially exclude sessions and/or participants with large head movements. However, none of the participants moved >3 mm during each

block. With the exception of participants 3 and 12 which were excluded for MRI artifacts not allowing respectively the coregistration and the anatomical reconstruction, all remaining 16 participants were considered for further analysis.

MarsAtlas-based cortical source model

To perform cortical reconstruction we used the FreeSurfer[1] (Fischl, 2012) toolbox, then to build surface meshes and perform parcelization we used the BrainVISA[2] (Cointepas et al., 2001) toolbox. The parcelization was performed following the MarsAtlas (Auzias et al., 2016) anatomical atlas, thus we obtained a total of 82 cortical parcels (41 each hemisphere). Then we used the BV2MNE toolbox[3], a python library developed in our team based on MNE-python[4] (Gramfort, 2013), able to transform the 3D spatial coordinates of the BrainVISA meshes back to MNI space for MNE compatibility, and to compute the Boundary Element Model (BEM), the source space, and the forward model. These three elements are needed for the power estimation at the source level, which will be discussed in the next paragraph. We performed the coregistration using the ‘mne coreg’ interface.

Single-trial High-Gamma Activity (HGA)

Preprocessing and artefact rejection

To analyse neurophysiological data we used a procedure similar to the one described in (Brovelli et al., 2015). All the following analyses were performed using the MNE toolbox (Gramfort, 2013). Raw MEG signals passed a visual inspection to check recording quality, two defective MEG sensors were excluded from the analysis for all participants. MEG signals were high pass filtered to 1Hz, low-pass filtered to 250 Hz, notch filtered in multiples of 50Hz and segmented into epochs aligned on outcome presentation (win/lose face). In the same way, a baseline epoch was computed from the 10 sec recording at the beginning of each scenario. We performed an Independent Component Analysis (ICA) taking the 95% of the whole explained variance in order to detect and reject cardiac, eye-blink and oculomotor associated

[1] https://surfer.nmr.mgh.harvard.edu/
[2] https://brainvisa.info/web/
[3] https://github.com/brainets/bv2mne
[4] https://mne.tools/stable/index.html

artifacts. Artifact rejection was performed semi automatically, at first we performed a visual inspection of the epochs' time series, then we used the autoreject python library (Jas et al., 2017) that uses machine learning and k-fold cross-validation methods to detect and reject bad epochs from further analysis.

Single-trial HGA in MarsAtlas

Spectral density estimation was performed using a multitaper method based on Discrete Prolate Spheroidal Sequences (DPSSs or Slepian Tapers; (Percival and Walden, 1993; Mitra and Pesaran, 1999)). We focused on HGA because it is well known to be a good neurophysiological marker for local mesoscopic event related activity (von Stein and Sarnthein, 2000; Ray and Maunsell, 2011; Buzsáki and Wang, 2012), and involved in higher cognitive processing (Scherberger et al., 2005; Gaona et al., 2011). To estimate the power of the high gamma band (from 60 to 120 Hz), MEG time series were multiplied by k orthogonal tapers (k = 11) (0.2 s in duration and 60 Hz of frequency resolution, each stepped every 0.005 s), centered at 90 Hz, and Fourier-transformed. Complex-valued estimates of spectral measures, including cross-spectral density matrices, were computed at the sensor level for each trial n, time t, and taper k.

In MEG we are interested in estimating the power of a signal at the level of virtual sources (dipoles) placed on the surface of the participants' 3D brain model. In order to pass from the sensor space to the source space, a forward model is needed. The forward model combines geometrical relations between sensors and sources with the BEM, which is a volume conduction model. For each participant, we generated a BEM using a single-shell model constructed from the segmentation of the cortical tissue obtained from individual MRI scans (Nolte, 2003). Those spatial and physical information were used to derive single-participant forward models.

We used adaptive linear spatial filtering (Veen et al., 1997) to estimate the power at the source level (inverse model). We used the Dynamical Imaging of Coherent Sources (DICS) method, a beam-forming algorithm for the tomographic mapping in the frequency domain (Gross et al., 2001), which is well suited for the study of neural oscillatory responses based on single-trial source estimates of band-limited MEG signals. At each source location, DICS algorithm uses a spatial filter that passes activity from this location with unit gain while maximally suppressing any other

activity. The spatial filters were computed on all trials for each time point and session and then applied to single-trial MEG data.

Once the single-trial high-gamma power at each source location was estimated both for the outcome aligned activity and for the baseline activity, we normalized the single-band power computing the relative change respect to the baseline defined as:

$$X_{source}^{n,\,t} = \frac{x_{source}^{n,\,t} - \frac{\sum_t b_{source}^{n,\,t}}{t}}{\frac{\sum_t b_{source}^{n,\,t}}{t}} \quad (1)$$

Where $X_{source}^{n,\,t}$ and $x_{source}^{n,\,t}$ corresponds respectively to the normalized and non-normalized power source estimate for each trial (n) and time point (t), and $\frac{\sum_t b_{source}^{n,\,t}}{t}$ correspond to the power source estimate of the baseline averaged over time.

Finally, we averaged the normalized source's power over the previously defined MarsAtlas parcel to obtain a single power time course for each region of interest (ROI). With this method we obtained for each participant a matrix describing the single-trial time course of HGA sampled at 200 Hz of 82 cortical brain regions (**Figure 3**).

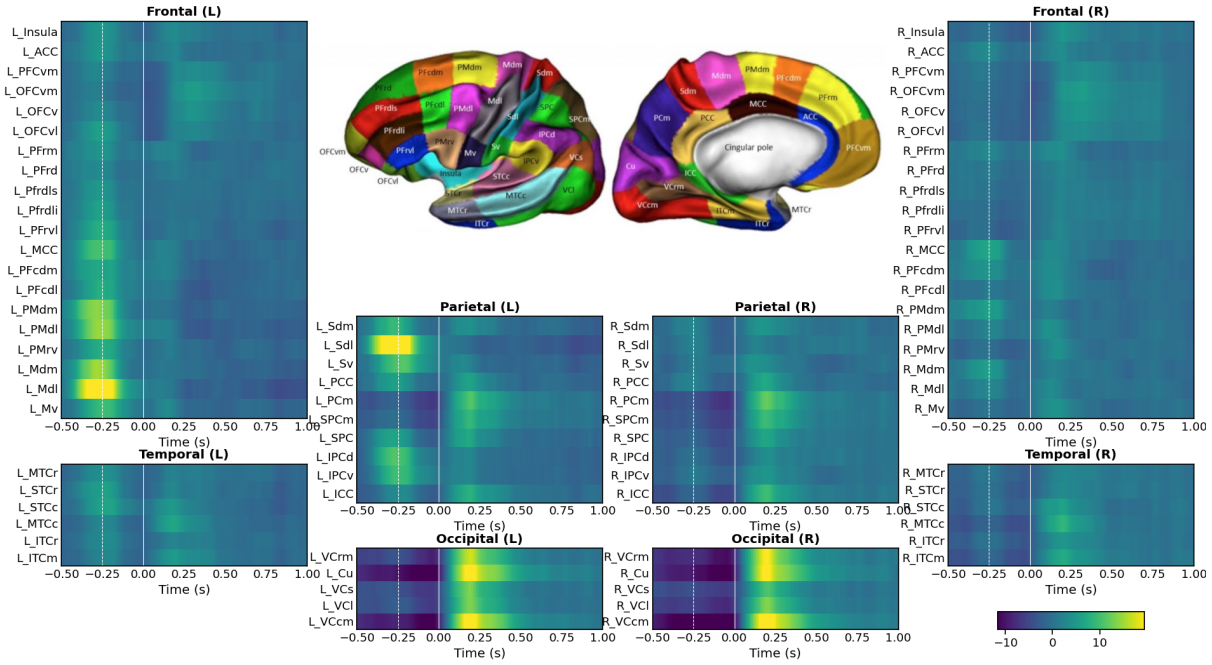


Figure 3. *MarsAtlas parcelization and average HGA. MarsAtlas is an anatomical atlas that comprehends a total of 82 cortical regions (41 each hemisphere). Here we showed the the average HGA computed across all the participants and all the trials, the figure clearly shows an event related potential interesting the motor, premotor, and sensorial cortices aligned with the action execution (-0.25 msec), and the visual activity triggered by the outcome delivery (0.0 msec).*

Bayesian ideal observer model of causal learning

In psychology, the contingency value ΔP is computed as the difference between two conditional probabilities: P(O|A), that is the probability of obtaining a positive outcome when the action is performed, and P(O|¬A) that is the probability of obtaining a positive outcome in absence of the action. Those two probabilities are independent, and they can be separately considered as Bernoulli distributions because of the binary nature of the outcome (0 = negative outcome; 1 = positive outcome). Thus, if we call the outcome result xi and its associated probability θ we can write the probability mass function as:

$$p(xi|\theta) = \begin{cases} \theta & \rightarrow \ if\ xi{=}1 \\ 1{-}\theta & \rightarrow \ if\ xi{=}0 \end{cases} \quad (2)$$

That can be written also as:

$$p(xi|\theta) = \theta^{xi}(1-\theta)^{1-xi} \text{ for } xi \in \{0,\ 1\} \quad (3)$$

Equation 3 describes the probability function for a single outcome event, but if we want to consider the whole sequence of equally likely outcomes obtained by the sequence of independent trials, expressed as a vector D of length n, we should rewrite this equation as:

$$p(D|\theta) = \prod_n p(xi|\theta) \tag{4}$$

Given that we are considering equally likely independent trials, this equation can be written as a binomial experiment, using the binomial coefficient notation. Thus, given n number of trials, k number of positive outcomes and $n - k$ number of negative outcomes, we can write:

$$\prod_n p(xi|\theta) = \binom{n}{k} (\theta^k (1 - \theta)^{n-k}) \tag{5}$$

$$= (n!/(k!\,(n - k)!))(\theta^k (1 - \theta)^{n-k}) \tag{6}$$

We are now interested in finding the distribution θ able to describe the data D. Since a binomial distribution describes the distribution of the outcomes but not the distribution of the trials' probabilities, we used Bayes rule using the binomial distribution for likelihood and a beta distribution ($B(\alpha, \beta)$) as conjugate prior. The product of the two generates a posterior beta distribution able to describe the distributions of the probabilities associated to the outcome observing the outcomes' results:

$$p(\theta|D) = B(k + \alpha;\, n - k + \beta) \tag{7}$$

As we can see, **Equation 7** is able to describe the update of beliefs depending on discrete states of the world, that in this case corresponds to the sequence of received outcomes, and on the total number of accumulated evidence, acting like an optimal Bayesian observer. The variables α and β can be considered as prior beliefs influencing the skewness and the shape of the beta distribution. We fixed those two values to 1.1 to give a symmetrical and constant prior belief that has an influence especially on the early trials.

To simplify the comprehension of how the model works we can see it in a frequentist way, as shown in **Figure 4A**. Having two possible actions with their independent outcome probability, we can define two distinct beta models:

$$p(\theta|D_A) = B(f(O|A) + \alpha;\ f(\neg O|A) + \beta) \quad \text{and}$$
$$p(\theta|D_{\neg A}) = B(f(O|\neg A) + \alpha;\ f(\neg O|\neg A) + \beta) \tag{8}$$

At each trial, when an action (play / no play) is performed and an outcome (win / lose) is received, only one of the four variables among the two models is updated (meaning consequently that only one of the models, the one corresponding to the chosen action, is updated). Thus, for each trial we can compute relevant behavioral variables associated to the participants behavior, for example we can compute P(O|A) and P(O|A) taking the mean of the two distributions, and then use those values to compute the updating belief about ΔP (**Figure 4B**):

$$\Delta P = \frac{f(O|A)}{f(O|A)+f(\neg O|A)} - \frac{f(O|\neg A)}{f(O|\neg A)+f(\neg O|\neg A)} \tag{9}$$

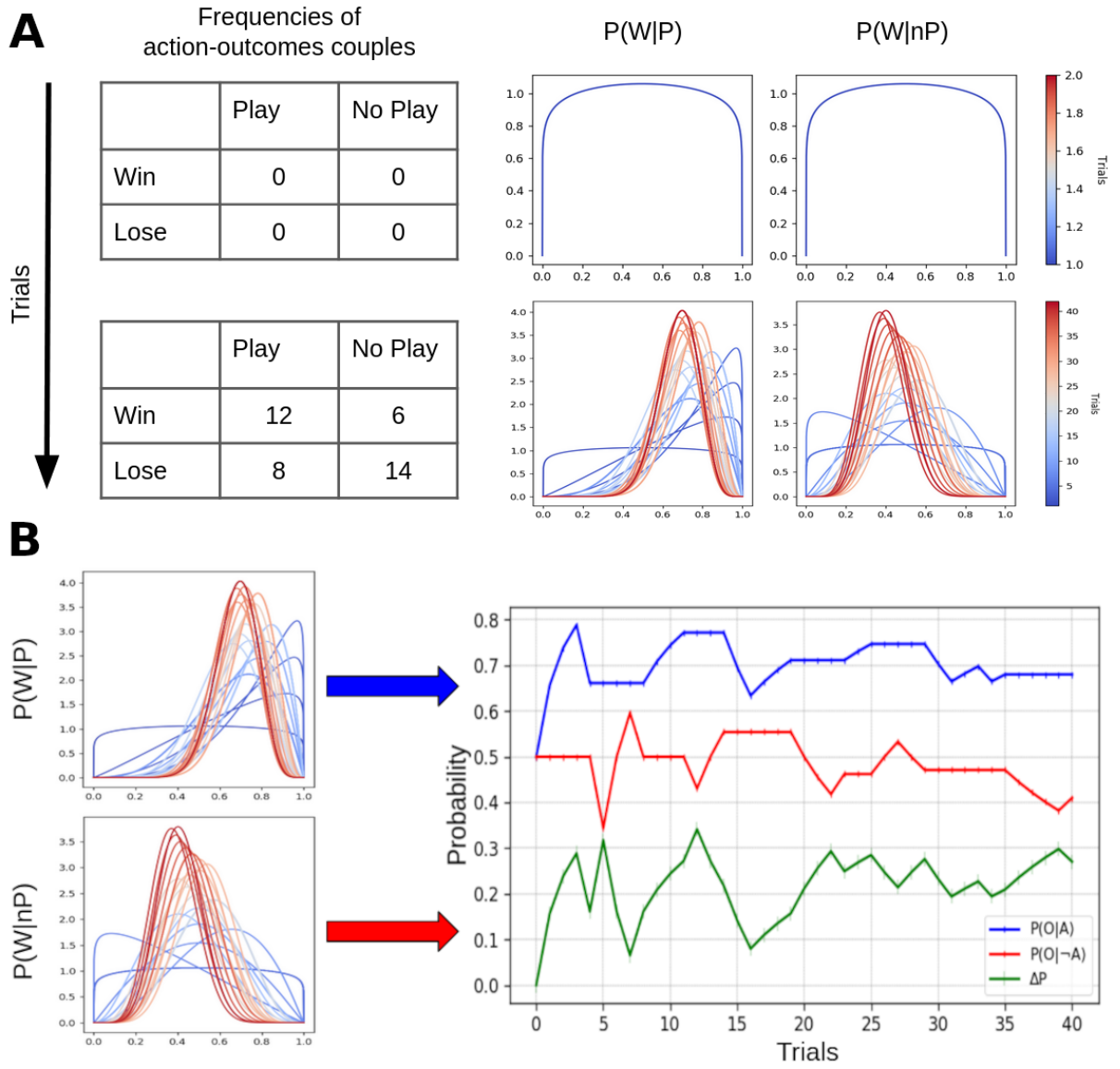


Figure 4. A) *Trial-by-trial evolution of the two contingency models. At the beginning both of the models' probability distributions are centered at p=0.5, which represents the probability of obtaining the outcome considering the past relative actions. At each trial one of the two models is updated according to the chosen action. The result of the outcome establishes the skewness of the probability distribution function.*
B) *At each trial we computed the average of both beta distributions, corresponding to the evolution of P(O|A) and P(O|¬A), and we used these two values to compute the ΔP.*

Model-based analysis of cortical HGA

Model-based information theoretical analysis

We used information-theoretic metrics to quantify the statistical dependency between single trial HGA and the model-based behavioral variables computed with the beta

model. Information-based measures quantify how much the neural activity of a single brain region explains a variable of the task. To this end, we computed the mutual information (MI); as a reminder, mutual information is defined as:

$$I(X;Y) = H(X) - H(X|Y) \quad (10)$$

In this equation the variables X and Y represent the HGA power and the behavioral variables, respectively. $H(X)$ is the entropy of X, and $H(X|Y)$ is the conditional entropy of X given Y. In particular, here we used Gaussian-Copula Mutual Information (GCMI) (Ince et al., 2017), that is a semi-parametric binning-free technique to calculate MI, in order to overcome some difficulties linked to the use of its classical version. Indeed MI requires sampling the full joint distribution of the two considered variables, making it difficult to estimate in case of limited amount of data, while the GCMI exploits the fact that the MI does not depend on the marginal distributions of the variables, but only on the copula function which encapsulates their statistical dependency. The GCMI results being a robust rank-based approach that allows to detect any type of relation as long as this relation is roughly monotone.

Statistical analysis

For the statistical inferences, we used a group-level approach based on non-parametric permutations and encompassing non-negative measures of information (Combrisson et al., 2021) implemented in the Frites[5] Python software. We used a random effect (RFX) to take into account the inter-subject variability, at the cost of needing a slightly larger dataset to achieve reliable statistical inferences. In this approach the MI between the neurophysiological signal and the behavioral regressor is computed across trials for each participant separately, at each time point and brain region. To sample the distribution of MI attainable by chance, we computed the MI between the brain data and a randomly shuffled version of the behavioral variable (Combrisson and Jerbi, 2015). This procedure was then repeated 1000 times. Thus, we took the mean of the MI values computed on the permutations, and used this mean(MI) to perform a one sample t-test across all the participants' MI values obtained both from original and permuted data. We then used a cluster-based

[5] https://github.com/brainets/frites

approach to assess whether the size of the estimated t-values significantly differs from its distribution. The cluster forming threshold was defined as the 95th percentile of the distribution of t-values. We used this cluster forming threshold to identify the cluster mass of t-values on both original and permuted data. Finally, to correct for multiple comparisons across both time and space, we build a distribution made of the 1000 largest clusters estimated on the permuted data. The final corrected p-values were inferred as the proportion of permutations exceeding the t-values.

Results

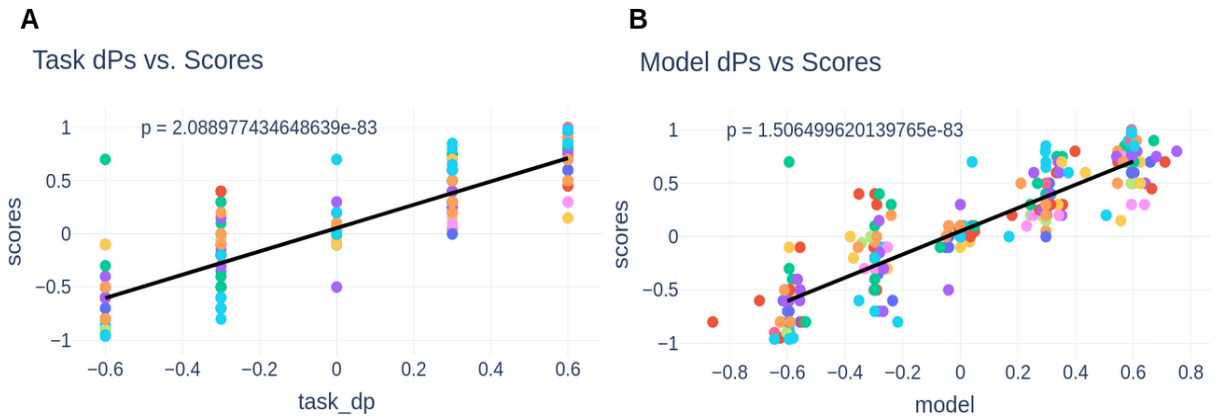


Figure 5. A) *Linear regression between the task associated ΔP values and the participants' scores.*
B) *Linear regression between the ΔP values computed by the Bayesian model and the participants' scores.*

Learning causal relations

Our first objective was to assess if participants correctly performed the task showing to be able to learn contingency values. At the end of each learning scenario, we asked participants to report the causal score they wanted to attribute to the evaluated player. The score could have been expressed by a number in an interval between -100 (negative causal perception) and 100 (positive causal perception), where the 0 represented an absence of causal relation. We divided these scores by 100 in order to rescale them between -1 and 1 and we performed a linear regression between them and the ideal ΔP values that we used in the task. As we can see in **Figure 5A**, we found a very strong significant (p≃2.089*10^-83) positive correlation between participants' causal perception and the ΔP values associated with the task, meaning that they were able to correctly estimate the hidden probabilities. Moreover, as shown in **Figure 6A**, the estimated values obtained by the linear regression, represented by the black line, are very close to the values that we would obtain if the participants were able to perfectly estimate the causal score, represented by the red line. Interestingly, participants' scores seem to be 'optimistic', in the sense that they are increasingly higher when they try to infer higher values of ΔP.

Model performance

We built a behavioral model based on an ideal Bayesian observer. This allowed us to obtain a trial-by-trial description of how the cognitive representation of the task related probabilities evolves during learning. To assess if the model is actually able to capture the participants' ability to encode causality, we performed a linear regression analysis between the last values of ΔP computed by the model at the end of each scenario and the score reported by participants normalized between -1 and 1. In **Figure 5B**, we can observe a significant ($p \approx 1.506*10^{-83}$) positive correlation between the causal effect computed by the beta model and the one reported by the participants at the end of each scenario. This means that the model can actually give a good indication of how the participants build their final representation of the contingency values associated with the task when observing the actions-outcomes succession. Moreover, in **Figure 6B** we plotted the variance of the final scores obtained by the model in function of the task ΔP values and the linear regression between those two variables (black line) in comparison with the perfect estimation of the task associated ΔP values (the red line). We can see just one of the two lines because they are almost overlapping, suggesting that the model truly act like an optimal observer.

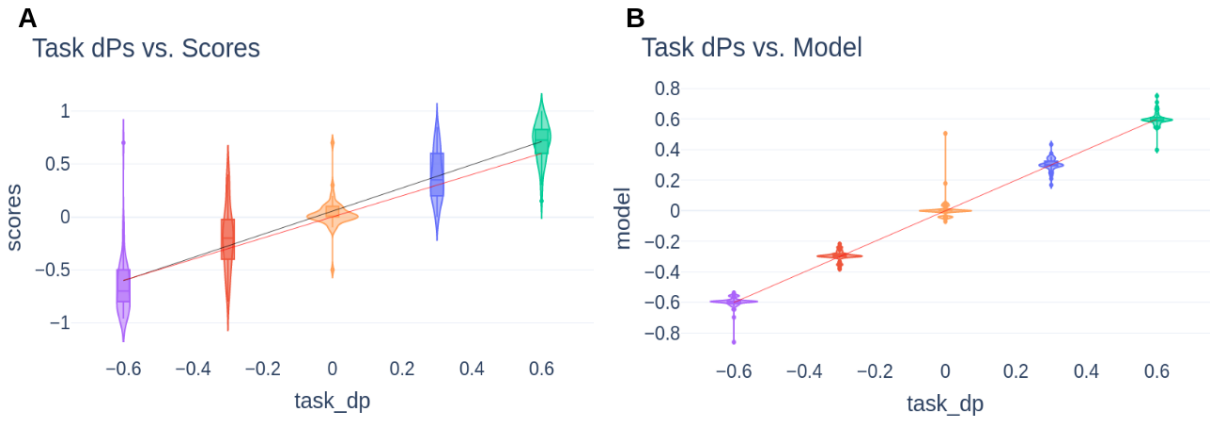


Figure 6. A) *Variance of the participants' scores relative to the task related ΔP values. The black line represents the linear regression computed in Figure 5, while the red line represents the ideal regression if the participants' were able to perfectly infer the ΔP values.*
B) *Variance of the model ΔP relative to the task related ΔP values. The black line represents the linear regression between the two variables, and is not visible because it is almost superposed to the red line representing the ideal regression if the model were able to perfectly predict the ΔP values.*

Neural correlates of instrumental contingency learning

We computed the trial-by-trial ΔP values with the Bayesian model and the single trial HGA of each participant, aligned on the outcome delivery. Then we estimated the MI between these two variables along the time series, and we performed statistics following the previously explained RFX protocol. We found a substantial increment of MI in the post-outcome period, in four prefrontal regions of the right hemisphere: the ventral, ventrolateral and ventromedial orbitofrontal cortices (OFCv, OFCvl and OFCvm respectively) and the rostral dorsolateral inferior prefrontal cortex (Pfrdli). However, after the statistic analysis only one of these regions resulted significant, that is the OFCv (**Figure 7**, p≈0.006) peaking around 0.5 seconds after the outcome, while the Pfrdli and the OFCvl were slightly higher the significance threshold (p≈0.053 and p≈0.074 respectively), and the OFCvm was well above the significance threshold (p≈0.193).

Neural correlates of actions-outcome probabilities

We followed the same pipeline used for the ΔP to find the neurophysiological correlates to the actions-outcome contingency values. In this case we computed from

the model two regressors: the probability of winning when the new player plays the match, or P(O|A), and the probability of winning when the new player doesn't play, or P(O|¬A). Additionally, we computed a regressor based on the task-related probability of winning according to the chosen action, that we called P(O|C). As shown in **Figure 7**, we found significant clusters of MI in seven different brain regions of the right hemisphere associated with the P(O|A). In the frontal lobe we found a long significant cluster interesting the insular cortex (IC, p≃0.001) from slightly before the outcome presentation until around 0.38 seconds, together with the post-outcome activity of the OFCvl (p≃0.012), the rostral ventrolateral prefrontal cortex (PFrvl, p≃0.038), and the rostroventral premotor cortex (PMrv, p≃0.001). In the parietal lobe we found significant MI clusters soon after the outcome presentation in the superior parietal cortex (SPC, p≃0.012) and in the dorsal inferior parietal cortex (IPCd, p≃0.038). In the temporal lobe, we found a cluster interesting the caudal superior temporal cortex (STCc, p≃0.017) from -0.04 to 0.21 seconds respect to the outcome. Surprisingly, we observed an increment of the MI in the left rostral dorsal and medial prefrontal cortices (PFrd and PFrm respectively, result not shown in figures), but no significant cluster were detected for any regions, responding to the P(O|¬A). Regarding P(O|C), we found 4 significant clusters in as many brain regions across both the hemispheres, two aligned on the action and two on the outcome. In the right frontal lobe, a post-outcome significant cluster was detected in the IC (p≃0.018), while in the OFCvl we found a first significant cluster aligned with the action (p≃0.018), and a second cluster above the significance threshold aligned with the outcome (p≃0.105). In the left temporal lobe, a post-outcome significant cluster was detected in the rostral superior temporal cortex (STCr, p≃0.048), and a significant cluster aligned with the action was found in the rostral inferior temporal cortex (ITCr, p≃0.046).

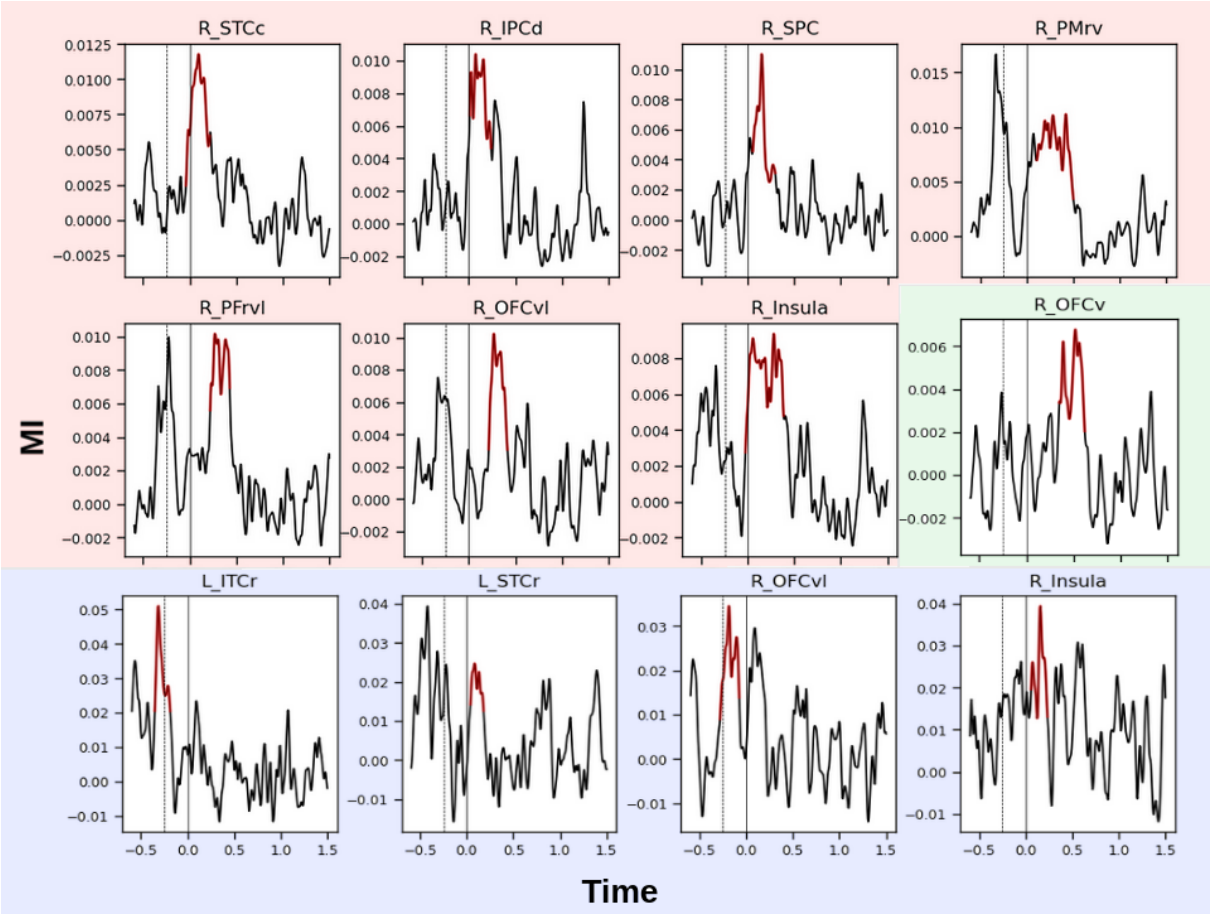


Figure 7. *Neural dynamic related to behavioral variables. Red lines represent significant clusters. On green background: MI with the ΔP values. On red background: MI with the P(O|A). On blue background: MI with the P(O|C).*

Behavior	ROI	Hemisphere	Full cortex name	Event	p-value
ΔP	OFCv	right	ventral orbitofrontal c.	outcome	p≃0.006
P(O|A)	IC	right	insular c.	outcome	p≃0.001
	OFCvl	right	ventrolateral orbitofrontal c.	outcome	p≃0.012
	PFrvl	right	rostral ventrolateral prefrontal c.	outcome	p≃0.038
	PMrv	right	rostroventral premotor c.	outcome	p≃0.001
	SPC	right	superior parietal c.	outcome	p≃0.012
	IPCd	right	dorsal inferior parietal c.	outcome	p≃0.038
	STCc	right	caudal superior temporal c.	outcome	p≃0.017
P(O|C)	IC	right	insular c.	outcome	p≃0.018
	STCr	left	rostral superior temporal c.	outcome	p≃0.048
	OFCvl	right	ventrolateral orbito frontal c.	action	p≃0.018
	ITCr	left	rostral inferior temporal c.	action	p≃0.046

Table 1. *Summary of the ROIs significantly correlating with the evolution of contingency (dP) and conditional action-outcome probabilities (P(O|A) and P(O|C)), and their cluster-based p-values.*

Discussion

In this study we investigated the temporal neural dynamics linked to causal learning. We asked human participants to perform a task in which they had to maximise their knowledge about the hidden contingencies of the task in a MEG machine. In order to model the trial-by-trial evolution of the contingency value and the relative probability of outcome given the chosen action we used an optimal observer Bayesian model based on a beta distribution. Finally we used MI and cluster based statistics to find significant relations between the time resolved high-gamma activity and the modeled behavioral variables.

Our results suggest a deep engagement of frontal, and especially prefrontal and orbitofrontal, cortical areas in encoding relevant aspects of causal learning, such as the contingency value (ΔP), the probability of the positive outcomes relative to the action 'play' (P(O|A)) and the task related probabilities of positive outcome given the chosen action (P(O|C)).

Participants' and model performances

The task that we proposed to participants is quite complex and requires more computational effort to be accomplished in comparison to classical contingency learning tasks. One of the differences is that the participant is not called to choose between performing an action and not performing it, but rather on choosing one action or another. The taken decision is then transferred to a middle agent (the player under evaluation) that is then supposed to execute (or not) the action. Moreover the goal of the task is less explicit, as we ask the participants to maximize their knowledge about the performance of the player under evaluation, and not, for example, to maximise the number of achieved positive outcomes.

Nonetheless, participants were able to learn contingency values and give an approximative final correct estimate of the causal scores. Most of them equally explored both of the possible actions. A common strategy was to change the chosen action each 3-5 trials. Questioning the participants after the task execution, some of them reported that sometimes they noticed late in the execution of a scenario that the action position was swapped. That can be the reason for some rare outliers that

emerged by the behavioral analysis. The model was able to reproduce participants' performances. The variance across the final modelled ΔP values, given by the differences in the sequence of actions performed by the participants, is lower with respect to the variance of the participants' scores. So far, the model is not able to explain the across participants variance. In the future we would like to take in consideration the behavioral variance in our model, fitting individually the initial prior probability (Lu et al., 2008). Another interesting model parameter to study would be the quantity of the post-outcome update increment, that can be related to the individual causal power perception of the actions (Cheng, 1997; Buehner et al., 2003).

Dynamic of the OFC in causal learning

Our results are in line with the literature, showing a prominent role of the OFC and in particular of its right-side rostro-ventral part in encoding information about the outcome identity and in discriminating the differences in outcome values. Despite most of the literature implicates the OFC in the encoding of the stimulus-outcome associations, for example in response to the presentation of a cue signaling a reward (Salzman et al., 2007; Salzman and Fusi, 2010; Howard et al., 2015), we should consider that in instrumental learning, in order to establish the relation between the stimulus and its outcome, an agent should be able to link the information about actions in a stimulus-action-outcome association (O'Doherty, 2007). Our results indicate that OFC can be sensible to the action value and that it can play a role in building a cognitive representation of the actions-outcomes probabilistic associations, indeed its implication in encoding the contingency value ΔP implies the knowledge of the conditional probabilities of the outcome given the action (Cheng, 1997; Hagmayer and Waldmann, 2007; Tanaka et al., 2008). In a fMRI study (Valentin et al., 2007) conducted on human participants performing an outcome devaluation task, the results suggested that the OFC is able to represent actions-outcomes information, showing a different activation profile for valued and devalued actions. This result is also in line with animals' studies performed on rats showed that prefrontal cortex and dorso-medial striatum are important to learn actions-outcomes association during goal directed learning (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Killcross, 2003). Moreover, the fact that OFC activity responds to ΔP and P(W|P)

after outcome presentation, and responds to P(W|C) after the action executions, can highlight its role both in acquisition and update of the actions-outcomes association and in outcome prediction.

Role of the PFC in encoding contingency

As the OFC, the prefrontal cortex (PFC) is implied in encoding outcome values. From our results we can see that the lateral rostro-ventral prefrontal cortex (PFrvl) participates in encoding positive outcome values but only if associated to the action 'play' and not to any chosen action. The ventral prefrontal cortex is known to mediate attentional processes and to encode stimulus salience (Asplund et al., 2010; Walther et al., 2011). Thus, we question whether this observed effect can be linked to an action dependent attentional mechanism, possibly derived by an unequal perception of the causal power attributable to the direct intervention of the agent ('play') rather than a random environmental variable ('no play'). Further investigations about the role of this region in the attentional processes linked to instrumental learning are needed.

Premotor cortex and insula

Also the premotor rostro-ventral cortex (PMrv) seems to be involved in encoding the P(W|P). This result is particularly challenging to discuss, as we would expect to find a modulation of the PMrv before the action selection, participating in action planning (Gremel and Costa, 2013), and not after the outcome presentation. This area has been defined as a relay from parietal to medial prefrontal cortices in visuomotor task (Viejo et al., 2015), but also in this case further investigations are needed.
Concerning the P(W|P) and the P(W|C), we found significant activation also in the insular cortex (IC) after receiving the outcome. The IC is known to participate in instrumental behavior in encoding incentive memories together with the amygdala (Parkes et al., 2015), in encoding the summed activity of potential outcomes (Liljeholm et al., 2013), and in retrieving outcome incentive values in order to guide the actions, but not in learning action-outcomes associations (Parkes et al., 2017). Thus, the activation of this region responding to these two behavioral regressors after the outcome presentation can be linked to the update of these values.

Parietal and temporal lobes

We observed an increment in MI in the parietal cortex in relation to ΔP values, however, statistical analysis showed that this increment is just below the significance threshold. Curiously, this effect turns out to be significant if we perform the analysis using the ΔP computed as log(P(W|P) / P(W|nP)) as behavioral regressors, while the effect found in the OFC is just below the significance threshold. Together, these result seems to indicate a role of the parietal cortex in encoding contingency values, also according to previous literature showing parietal cortex tracks contingency values computed both as ΔP and as the Jensen-Shannon divergence between the probabilities of the outcome conditioned on different actions (Liljeholm et al., 2011, 2013).

Regarding the temporal lobe, its implication in instrumental learning is less understood in comparison to other regions, nonetheless its activity has been related to formation and updating inferences about optimal behavioral strategies (O'Doherty et al., 2017).

Conclusion

The aim of this study is to assess the temporal dynamics of the contribution of different cortical brain regions during contingency acquisition in humans. To do so we instructed human participants in performing a goal-directed causal learning task under MEG recording. Then we built a Bayesian optimal observer model to perform single-trial model-based analysis. We found a prominent role of the OFC and the rostral PFC in encoding relevant behavioral correlates, such as the ΔP and the P(O|A). Despite most of the results presented here confirming previous findings, we believe that the neurophysiological correlates of goal directed causal learning needs a deeper investigation. That can be achieved going in two directions: I) improving the Bayesian model enhancing the fitting of the single subject behavior through the fine tuning of relevant parameters, and II) investigating the oscillatory activity of lower frequency bands. Indeed, in this work we focused specifically on the high-gamma band oscillatory activity. Moreover, It will be interesting to study the cortico-cortical interactions between pairs of brain regions forming functional networks supporting causal learning in time, using techniques such as Granger Causality (GC) or Partial Information Decomposition (PID).

Bibliography

- Allan, L.G. (1993). Human Contingency Judgments: Rule Based or Associative? Psychol. Bull. *114*, 435–448.
- Allan, L.G., and Jenkins, H.M. (1980). The judgment of contingency and the nature of the response alternatives. Can. J. Psychol. Can. Psychol. *34*, 1.
- Allan, L.G., Hannah, S.D., Crump, M.J.C., and Siegel, S. (2008). The psychophysics of contingency assessment. J. Exp. Psychol. Gen. *137*, 226–243.
- Asplund, C.L., Todd, J.J., Snyder, A.P., and Marois, R. (2010). A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. Nat. Neurosci. *13*, 507–512.
- Auzias, G., Coulon, O., and Brovelli, A. (2016). MarsAtlas: A cortical parcellation atlas for functional mapping. Hum. Brain Mapp. *37*, 1573–1592.
- Balleine, B.W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology *37*, 407–419.
- Blaisdell, A.P. (2006). Causal Reasoning in Rats. Science *311*, 1020–1022.
- Brovelli, A., Lachaux, J.-P., Kahane, P., and Boussaoud, D. (2005). High gamma frequency oscillatory activity dissociates attention from intention in the human premotor cortex. NeuroImage *28*, 154–164.
- Brovelli, A., Chicharro, D., Badier, J.-M., Wang, H., and Jirsa, V. (2015). Characterization of Cortical Networks and Corticocortical Functional Connectivity Mediating Arbitrary Visuomotor Mapping. J. Neurosci. *35*, 12643–12658.
- Buehner, M.J., Cheng, P.W., and Clifford, D. (2003). From Covariation to Causation: A Test of the Assumption of Causal Power. J. Exp. Psychol. Learn. Mem. Cogn. *29*, 1119–1140.
- Buzsáki, G., and Wang, X.-J. (2012). Mechanisms of Gamma Oscillations. Annu. Rev. Neurosci. *35*, 203–225.
- Cheng, P.W. (1997). From covariation to causation: a causal power theory. Psychol. Rev. *104*, 367.
- Cointepas, Y., Mangin, J.-F., Garnero, L., Poline, J.-B., and Benali, H. (2001). BrainVISA: Software platform for visualization and analysis of multi-modality brain data. NeuroImage *13*, 98.
- Combrisson, E., and Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J. Neurosci. Methods *250*, 126–136.
- Combrisson, E., Allegra, M., Basanisi, R., Ince, R.A.A., Giordano, B., Bastin, J., and Brovelli, A. (2021). Group-level inference of information-based measures for the analyses of cognitive brain networks from neurophysiological data (Neuroscience).
- Corbit, L.H., and Balleine, B.W. (2003). The role of prelimbic cortex in instrumental conditioning. Behav. Brain Res. *146*, 145–157.
- Dickinson, A., and Balleine, B.W. (2000). Causal Cognition and Goal-Directed Action. Evol. Cogn. 185.
- Dolan, R.J., and Dayan, P. (2013). Goals and Habits in the Brain. Neuron *80*, 312–325.

- Fischl, B. (2012). FreeSurfer. NeuroImage *62*, 774–781.
- Gaona, C.M., Sharma, M., Freudenburg, Z.V., Breshears, J.D., Bundy, D.T., Roland, J., Barbour, D.L., Schalk, G., and Leuthardt, E.C. (2011). Nonuniform High-Gamma (60-500 Hz) Power Changes Dissociate Cognitive Task and Anatomy in Human Cortex. J. Neurosci. *31*, 2091–2100.
- Gramfort, A. (2013). MEG and EEG data analysis with MNE-Python. Front. Neurosci. *7*.
- Gremel, C.M., and Costa, R.M. (2013). Premotor cortex is critical for goal-directed actions. Front. Comput. Neurosci. *7*.
- Hagmayer, Y., and Waldmann, M.R. (2007). Inferences about unobserved causes in human contingency learning. Q. J. Exp. Psychol. *60*, 330–355.
- Hammond, L.J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. J. Exp. Anal. Behav. *34*, 297–304.
- Howard, J.D., Gottfried, J.A., Tobler, P.N., and Kahnt, T. (2015). Identity-specific coding of future rewards in the human orbitofrontal cortex. Proc. Natl. Acad. Sci. *112*, 5195–5200.
- Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., and Schyns, P.G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula: Gaussian Copula Mutual Information. Hum. Brain Mapp. *38*, 1541–1573.
- Jas, M., Engemann, D.A., Bekhti, Y., Raimondo, F., and Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. NeuroImage *159*, 417–429.
- Killcross, S. (2003). Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. Cereb. Cortex *13*, 400–408.
- Liljeholm, M. (2018). Instrumental Divergence and Goal-Directed Choice. In Goal-Directed Decision Making, (Elsevier), pp. 27–48.
- Liljeholm, M. (2021). Agency and goal-directed choice. Curr. Opin. Behav. Sci. *41*, 78–84.
- Liljeholm, M., Tricomi, E., O'Doherty, J.P., and Balleine, B.W. (2011). Neural Correlates of Instrumental Contingency Learning: Differential Effects of Action-Reward Conjunction and Disjunction. J. Neurosci. *31*, 2474–2480.
- Liljeholm, M., Wang, S., Zhang, J., and O'Doherty, J.P. (2013). Neural Correlates of the Divergence of Instrumental Probability Distributions. J. Neurosci. *33*, 12519–12527.
- Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *412*, 8.
- Lu, H., Yuille, A.L., Liljeholm, M., Cheng, P.W., and Holyoak, K.J. (2008). Bayesian generic priors for causal learning. Psychol. Rev. *115*, 955–984.
- Meyniel, F., Schlunegger, D., and Dehaene, S. (2015). The Sense of Confidence during Probabilistic Learning: A Normative Account. PLOS Comput. Biol. *11*, e1004305.
- Mitra, P.P., and Pesaran, B. (1999). Analysis of Dynamic Brain Imaging Data. Biophys. J. *76*, 691–708.
- Moore, J.W., Lagnado, D., Deal, D.C., and Haggard, P. (2009). Feelings of control: Contingency determines experience of action. Cognition *110*, 279–283.
- Morris, R.W., Dezfouli, A., Griffiths, K.R., Le Pelley, M.E., and Balleine, B.W. (2017). The algorithmic neuroanatomy of action-outcome learning (Neuroscience).

- Msetfi, R.M., Wade, C., and Murphy, R.A. (2013). Context and Time in Causal Learning: Contingency and Mood Dependent Effects. PLoS ONE *8*, e64063.
- Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. Phys. Med. Biol. *48*, 3637–3652.
- O'Doherty, J.P. (2007). Lights, Camembert, Action! The Role of Human Orbitofrontal Cortex in Encoding Stimuli, Rewards, and Choices. Ann. N. Y. Acad. Sci. *1121*, 254–272.
- O'Doherty, J.P., Cockburn, J., and Pauli, W.M. (2017). Learning, reward, and decision making. Annu. Rev. Psychol. *68*, 73–100.
- Parkes, S.L., Bradfield, L.A., and Balleine, B.W. (2015). Interaction of Insular Cortex and Ventral Striatum Mediates the Effect of Incentive Memory on Choice Between Goal-Directed Actions. J. Neurosci. *35*, 6464–6471.
- Parkes, S.L., Ravassard, P.M., Cerpa, J.-C., Wolff, M., Ferreira, G., and Coutureau, E. (2017). Insular and Ventrolateral Orbitofrontal Cortices Differentially Contribute to Goal-Directed Behavior in Rodents. Cereb. Cortex 1–13.
- Penn, D.C., and Povinelli, D.J. (2007). Causal Cognition in Human and Nonhuman Animals: A Comparative, Critical Review. Annu. Rev. Psychol. *58*, 97–118.
- Percival, D.B., and Walden, A.T. (1993). Spectral Analysis for Physical Applications (Cambridge University Press).
- Ray, S., and Maunsell, J.H.R. (2011). Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. PLoS Biol. *9*, e1000610.
- Salzman, C.D., and Fusi, S. (2010). Emotion, Cognition, and Mental State Representation in Amygdala and Prefrontal Cortex. Annu. Rev. Neurosci. *33*, 173–202.
- Salzman, C.D., Paton, J.J., Belova, M.A., and Morrison, S.E. (2007). Flexible Neural Representations of Value in the Primate Brain. Ann. N. Y. Acad. Sci. *1121*, 336–354.
- Scherberger, H., Jarvis, M.R., and Andersen, R.A. (2005). Cortical Local Field Potential Encodes Movement Intentions in the Posterior Parietal Cortex. Neuron *46*, 347–354.
- Shanks, D.R. (1985). Continuous monitoring of human contingency judgment across trials. Mem. Cognit. *13*, 158–167.
- Shanks, D.R., and Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. Mem. Cognit. *19*, 353–360.
- von Stein, A., and Sarnthein, J. (2000). Different frequencies for different scales of cortical integration: from local gamma to long range alphartheta synchronization. Int J Psychophysiol *38*, 301–313.
- Tanaka, S.C., Balleine, B.W., and O'Doherty, J.P. (2008). Calculating Consequences: Brain Systems That Encode the Causal Effects of Actions. J. Neurosci. *28*, 6750–6755.
- Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the Neural Substrates of Goal-Directed Learning in the Human Brain. J. Neurosci. *27*, 4019–4026.
- Viejo, G., Khamassi, M., Brovelli, A., and Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. Front. Behav.

Neurosci. *9*.
- Walther, S., Friederich, H.-C., Stippich, C., Weisbrod, M., and Kaiser, S. (2011). Response inhibition or salience detection in the right ventrolateral prefrontal cortex? NeuroReport *22*, 778–782.
- Wasserman, E.A., Chatlosh, D.L., and Neunaber, D.J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. Learn. Motiv. *14*, 406–432.
- Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005). The role of the dorsomedial striatum in instrumental conditioning: Striatum and instrumental conditioning. Eur. J. Neurosci. *22*, 513–523.

Section 5. Neuroinformatics, tools, Open Science

5.1 Team resources

During my PhD I wrote pipelines for future students in order to allow them to easily access team resources. A first pipeline is focused on the creation and organization of neurophysiological dataset on the INT's high performance computing cluster 'frioul'.

A second pipeline is focused on how to use BrainVISA and FreeSurfer to manage and analyse MRI data, in order to compute brain volumetric space and cortical surface of participants' MRI, to then apply the MarsAtlas parcelization.

A third pipeline is focused on the use of rsync, an informatic tool that we use to transfer data on and between clusters and local machines. All the pipelines are open and accessible at BraiNets' resources GitHub page (https://github.com/brainets/ressources).

5.2 Softwares development

In collaboration with David Menieur, I developed a python library called BV2MNE (https://github.com/brainets/bv2mne). This library acts like an interface between BrainVisa and MNE softwares. Indeed, it can access the cortical meshes generated by BrainVisa to transform them in a MNI space, a format that is compatible with MNE. Starting from the transformed meshes, BV2MNE computes the boundary element model (BEM) and a labelized source space that will be used to compute the forward and inverse model (i.e. the sensor signal reconstruction at the source level). Moreover, the library contains visualization functions able to show source

disposition and lablization in a 3D dynamic space, together with the cortical meshes and the BEM.

I also contributed in Etienne Combrisson python library 'Frites' (FRamework for Information Theoretical analysis of Electrophysiological data and Statistics; https://github.com/brainets/frites), especially in the testing part, and in the conversion of CPU function for the GPU use. Frites is a toolbox for assessing information-theoretic measures on human and animal neurophysiological data, to extract task-related cognitive brain dynamics and perform group-level statistics.

5.3 Open science

During my PhD I dedicated part of my time to the culture of open science, especially with the participation and organization of BrainHack events hosted in Marseille. BrainHack Global (https://brainhack.org/) is a community promoting the culture of open science all over the world. Through the organization of local events (both in person and on-line), BrainHack leads people from different fields with their own skills and ideas to meet up and join already existing projects or proposing their own. In 2019 I presented BV2MNE as a project in the BrainHack Marseille community. In 2020 I participated in the organization of BrainHack Marseille (https://brainhack-marseille.github.io/), and I proposed a project together with Etienne Combrisson on GPU porting of Frites' python library functions. The culture of open science is important to promote scientific collaborations in an inclusive environment, to improve the flow of information between domains of the same or different disciplines. It can raise scientists' awareness on the use and misuse of data, and of big datasets, that are so expensive to collect. Moreover it poses an accent on the importance of having common and reproducible good practices, to enhance research workflows. Last but not least, it is an occasion to meet other people and increase networking.

5.4 NeuroMatch Academy - Deep Learning

In summer 2021, I attended the one month NeuroMatch Academy summer school on ANN and deep learning. In the last years, ANN found vaste applications in several fields, indeed especially after the advent of Deep Neural Networks (DNNs) (Shrestha and Mahmood, 2019; Emmert-Streib et al., 2020) and Deep Learning (DL) (Goodfellow et al., 2016) almost all devices and softwares of everyday use extensively exploits this technology . DNN have become famous for their efficiency in solving different tasks, with more or less variations in their architectures and learning algorithms. For example, Convolutional Neural Networks (CNNs) (Gu et al., 2018) are mostly used to perform image processing, while Recurrent Neural Networks (RNNs) (Buesing et al., 2011; Lipton et al., 2015) are particularly suitable for natural language processing. A great innovation was represented by generative deep networks, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014), that are not only able to classify inputs, but also to produce them by stochastically sampling learned characteristics of the corresponding input from a continuous features' space. Among the limitations of DNNs we can indicate the needed long training time, and the fact that they are data-hungry. Whereas, the limitations of DNNs related to the study of the brain can be attributed to their biological implausibility, due to the use of artificial algorithms to improve learning and reduce learning time, as for example the very backpropagation. There are recent studies that propose to relate brain areas activity to DNNs' layers representations, using Representational Similarity Analysis (RSA) (Kriegeskorte, 2008).

Section 6. Discussion and Conclusion

Goal-directed behavior (GDB) is a flexible but computationally heavy decision making strategy that allows us to face novel problems in an adaptable way. It comprehends several non-trivial cognitive aspects, making it hard to model and to study in terms of underlying neural and network computations. At first, we need to have a mental representation of the goal, a desired future state to which we want to tend. Then, to achieve this goal, we should be able to represent the space of actions-outcome combinations in a cognitive map and to use and update it in order to plan our future responses, balancing our behavior in between exploring or exploiting actions after evaluating received outcomes. Moreover, if the rules of the context change, this cognitive map should be flexible enough to allow rapid behavioral adjustments. This complexity is supported by intricate brain networks involving neural computations at both cortical and subcortical levels, in particular prefrontal, orbitofrontal and parietal cortical regions, the hippocampus, and the whole striatum.

The aim of this thesis was to expand our current knowledge about neural and computational mechanisms that give rise to GDB through the analysis of three aspects:

1) **Neural network mechanisms for goal-directed learning:** Identify potential neural network mechanics able to learn the world model contextually to its use for planning and to learn such a world model in an autonomous fashion based on unsupervised learning processes, using a spiking neural-network architecture bridging planning as inference and brain-like mechanisms.

2) **Anatomo-functional gradients along the striatum for goal-directed learning:** Describe the role of the striatum in nonhuman primates in goal-directed learning, and in particular its activity in response to RPE; a relevant behavioral signal used to update the values of state-dependent actions in response to reward, through the use of a Q-learning model.
3) **Cortical circuits for causal goal-directed learning:** Outline the possible contributions of different brain areas in humans during the execution of a goal-directed causal learning task, relating their brain activity with trial-by-trial behavioral variables obtained through the use of an optimal agent Bayesian model.

In the next sections I will discuss and give a future perspective of each of these three aspects to then give a general possible future perspective related to this research topic.

6.1 A generative spiking neural-network model of goal-directed behaviour and one-step planning - Faced problems and solutions to GDB modelling

The spiking neural network described in **Section 2** exhibits two relevant features: its capacity to autonomously form neural internal representations (hidden causes) of the observations at different times and to activate them in sequence; and that the spiking sampling probability reflects the probability distributions expressed by the world model. Thus this model is not only able to build a representation of the world in an unsupervised way, but also to use it to simulate stimuli-actions-outcomes trajectories that can be used to plan goal-directed actions to achieve a desired state. These properties of the network arise from Spike-Timing Dependent Plasticity (STDP) unsupervised learning rule and the features of the model architecture,

together with the stochastic nature of the model. Altogether, they allow the model to 'imagine' different action-feedback sequences in correspondence to a stimulus.
The ensemble of these complex processes arising from this neural architecture highlights another property of the network, that is its emergent generativity. Emergent generativity is characterised by two relevant elements. The first element regards 'generativity' and involves the stochastic nature of spike sampling that allows the production of alternative patterns in correspondence to the same context. This means that the network is able to form new chains of consecutive spikes even if that sequence was never observed before, using the previously learned representations of the states. This process is important as the generation of alternative plausible patterns is at the core of search algorithms possibly employed by the brain. For example, generativity can support the search of different courses of action that might lead to a desired goal state starting from a given initial condition. In deep neural networks, generativity is often based on stochastic elements supporting the generation of novel plausible patterns, as it happens in Generative Adversarial Networks (GANs; (Goodfellow et al., 2014)) and Variational Autoencoders (VAEs; (Kingma and Welling, 2014)) able to generate new plausible input patterns by drawing sample patterns from prior probability distributions and then by transforming them through deterministic neural components trainable with supervised learning. The second important element of emergent generativity regards 'emergence' and involves the process for which in complex systems, such as the brain, the dynamical interaction of low-level elements can give rise to organised patterns at higher levels (Newman, 2011). In particular, in the brain, events involving spike neurons at a low level are amplified by neural mechanisms in order to generate patterns that encode content, such as perceptions, thoughts and actions, at a higher cognitive level.

The obtained results highlight the novelties of the architecture that we proposed with respect to other current models. A first novelty with respect to the previous

models implementing planning as inference based on brain-like mechanisms (Friedrich and Lengyel, 2016; Rueckert et al., 2016; Tanneberg et al., 2016) is that our architecture proposes an hypothesis on how organisms might learn the world model while using it for planning. This is a key challenge for planning, as recently highlighted in (Tschantz et al., 2020). The challenge is different from the exploration/exploitation issue in model-free models (Sutton and Barto, 1998), and requires arbitration mechanisms different from the classic ones used to balance goal-directed and habitual processes (Daw et al., 2005; Viejo et al., 2015).

A second novel feature that allowed the architecture to autonomously learn the world

model is the use of a Hidden Markov Model (HMM) having a relevant difference with respect to those used in other planning-as-inference spiking network models (Rueckert et al., 2016; Tanneberg et al., 2016). These models use a world model-based on a classic HMM reproducing possible sequences of states but not actions. Instead, the world model used here is based on a HMM that observes sequences of states and of actions. This allows the world model to directly select actions to perform; instead, previous models (Rueckert et al., 2016; Tanneberg et al., 2016) need an additional mechanism selecting actions on the basis of the state sequence produced by the world model. Moreover, for each environment state the world model can suggest the selection of actions that have a potential relevance in that context, rather than any action, along with the idea of affordance in cognitive science (Baldassarre et al., 2019).

A third and last novel feature that allowed the architecture to autonomously learn the world model is the explicit representation of the goal used to condition the probability distribution expressed by the world model. Previous state-of-the-art models (Rueckert et al., 2016; Tanneberg et al., 2016) combined goal, initial state, and environment conditions into a whole ‘context’ representation. Our representation of goals allows their manipulation independently of other

conditions, as shown by the model's capacity to successfully plan how to reach new goals on the basis of the experience that the world model acquired in other tasks.

Finally, we used the model to try to reproduce the behavioral performances and the reaction times of a previous study on human participants performing a goal-directed learning task (Brovelli et al., 2008). The architecture has reproduced the target behaviour, has furnished an explanation of the mechanisms possibly underlying it, and has proposed predictions testable in future empirical experiments.

Despite that we acknowledge that the model has various limitations that might be improved in future work. A first open issue concerns the generalisation from neurons firing at discrete times to neurons firing in continuous time. This might be done using the inhomogeneous Poisson process (Kappel et al., 2014). Although this would not drastically modify the theoretical contribution of the model, it might simplify a comparison with real data from the brain at a finer temporal level with respect to what is done here. A second issue to face would be to use other tasks with respect to the one considered here (Brovelli et al., 2008), to test its robustness and capacity to scale-up to more complex tasks. A relevant issue to face in future work concerns the new arbitration mechanism proposed with the model. Here, the entropy measure at the core of the arbitration mechanism is grounded on the probability distribution of neurons. However, it is now hardwired, and future work should thus aim to implement this process with neural mechanisms. Another improvement of the model might involve the full development of a habitual component, not included here because it is out of the scope of this work.

6.2 Beta oscillations in the monkey striatum encodes reward prediction error - The role of beta-band oscillations in striatal RPE signaling

In the study described in **Section 3**, we recorded Local Field Potentials' (LFPs) activity in the striatum of monkeys performing a free-choice probabilistic learning task. The task required monkeys to choose between three options for movement, each one associated with different reward probabilities. The aim of the study was to investigate modulations in striatal activity, using this probabilistic design in order to detect changes specific to the processing of actions' outcomes (i.e. rewards). In particular, we aimed at studying the encoding of reward prediction errors in the monkey striatum. This task is well suited for studying action selection guided by predictions about future events and comparisons of those predictions with actual outcomes which correspond to Reward Prediction Error (RPE), RPE is an error signal generated by midbrain dopaminergic neurons (Abler et al., 2006; Schultz, 2007, 2016a) crucial to modify our behavior in order to improve the predictions about possible environmental outcomes (O'Doherty et al., 2017), thus playing an essential role in GDB (Ressler, 2004; Keramati et al., 2011). The role of midbrain dopaminergic neurons in RPE encoding is well established (Abler et al., 2006; Schultz, 2007, 2016a). Animal electrophysiology and human neuroimaging have also found evidence of RPE-related activity in the striatum (Schultz, 2016b). It is considered as a crucial signal that adaptatively support GDB, contributing to the world model updating and action planning (Takikawa et al., 2002; Ressler, 2004; Izawa and Shadmehr, 2011; Keramati et al., 2011; Schultz, 2016b) .

Three main results about striatal functional organization emerge from this study:

(1) we observed a significantly different pattern of oscillation in the high beta - low gamma frequency band when contrasting rewarded trials with non-rewarded trials;

(2) we found a significant increment in mutual information between the beta band

oscillation and the RPE computed using a Q-learning algorithm fitted on monkeys' choice behavior during the task;
(3) we divided the data according to the recording sites in eighteen clusters respecting the striatal anatomical constraints, then we computed the MI between the data of each cluster and the respective RPE, and we observed a significant positive linear correlation between the rostro-caudal and dorso-ventral clusters distribution and the total amount in time of RPE related MI, which we refer to an anatomo-functional gradient

A key finding in our study is the role of the beta band oscillations in carrying information about RPE in the basal ganglia system. Beta band activity is historically linked to motor control. Indeed, oscillations in this band have been observed in the motor cortex, especially associated with specific movements, like for example precision movements (Feingold et al., 2015; Khanna and Carmena, 2017). Moreover, after the advent of deep brain stimulation, data recorded from the STN of Parkinson's disease (PD) patients showed that an abnormal increase in beta oscillatory activity is associated to a lack of dopaminergic signaling, leading to Parkinsonian symptoms, and that a stimulation of the STN can interrupt this oscillatory beta activity and block PD motor symptoms (Kühn et al., 2004; Holt et al., 2019). Lately, some studies have reconsidered the role of beta band in the striatum, showing that it can be related to other important behavioral features and to reward value (Leventhal et al., 2012; Münte et al., 2017; Schwerdt et al., 2020). To our knowledge, this is the first report to demonstrate that outcome processing is an important variable influencing striatal beta activity in the nonhuman primate.

As we pointed out earlier, the RPE signal is essential to modulate several aspects of behavior. Thus, one can expect that this signal should be integrated within different domains of the striatum in order to participate in various functional processes involving limbic, associative, and motor cortico-striatal circuits (Oya et al., 2005;

Gläscher et al., 2010; Mestres-Missé et al., 2012; Vogelsang and D'Esposito, 2018). It is well known that dopamine is responsible for RPE signaling in the striatum (Abler et al., 2006; Schultz, 2016b). Furthermore, we know that around 95% of striatal neurons in monkeys are gabaergic medium spiny neurons (MSN), long projection neurons innervating pallidal and nigral areas, while the remaining 5% (this ratio follows interspecific changes) is composed of cholinergic interneurons (Lecumberri et al., 2017) spreading their axons across striatum, connecting NAcc, putamen (Put), and caudate (Cau) nuclei (Assous and Tepper, 2019). Further studies about the role of the striatal interneurons and the internal transmission of information can help us understand how RPE signal propagates across the striatum.

Another major finding was that the information about RPE is present, to varying degrees, in all territories of the striatum, forming a fading gradient stronger toward the rostro-ventral striatum and weaker toward its caudo-dorsal part. This result is in line with other studies, in which striatal circuitry is able to establish different functional gradients, spanning from the dopaminergic signaling to the cognitive control (Mestres-Missé et al., 2012; Vogelsang and D'Esposito, 2018; Alberquilla et al., 2020; Han et al., 2021), determined by the cortico-striatal and striato-thalamic loops. At the same time, this hypothesis casts a new light on the idea that basal ganglia, and especially striatum, can be roughly divided in functional regions, participating in limbic, associative or motor functions. Indeed, the idea of a neat functional division that was established, especially in other basal ganglia's structures like the STN (Eisinger et al., 2018), is lately going through a review (Alkemade and Forstmann, 2014; Eisinger et al., 2019). It is less astonishing to support the idea of gradients if we take in consideration the behavioral salience of the RPE. Indeed, also if a well structured connectivity is needed to transmit precise signals, the information contained in those can participate in other behavioral functions. RPE is needed to update the inner model of action values in response to a particular state, and those values should be retained in short term memory in order

to plan future actions in a goal-directed way. Our results are in line with the idea that the RPE is an important signal affecting several aspects of the behavior, and that for this reason it should propagate in limbic, associative, and motor cortico-striatal circuits. Understanding how this gradient rises from striatal connectivity remains to be elucidated.

6.3 Dynamics of human cortical circuits mediating goal-directed causal learning - High-gamma activity in human prefrontal cortex reflects relevant behavioral aspects of goal-directed causal learning

In the study described in **Section 4**, we investigated the functional role of prefrontal cortical areas and their implications in goal-directed causal learning. We asked human participants to perform a task in which they had to maximise their knowledge about the hidden contingencies of the task and to report us the supposed causal score at the end of each recording block, while being recorded in a MEG machine. In order to model the trial-by-trial evolution of the contingency value and the relative probability of outcome given the chosen action we used an optimal agent Bayesian model-based on a beta distribution (i.e., the ideal observer model). Finally we used MI to find significant relations between the estimated high-gamma activity in time and the modeled task-related behavioral variables.

Our results suggest a deep engagement of frontal, and especially prefrontal and orbitofrontal, cortical areas in encoding relevant aspects of causal learning, such as the contingency value (ΔP), the probability of the positive outcomes relative to the action ‘play’ ($P(W|P)$) and the task related probabilities of positive outcome given the chosen action ($P(W|C)$).

Our results are in line with the literature, showing a prominent role of the OFC and in particular of its right-side rostro-ventral part in encoding information about the outcome identity and in discriminating the differences in outcome values. Surprisingly, our results indicate that OFC can be sensible to the action value and that it can play a role in building a cognitive representation of the actions-outcomes probabilistic associations, indeed its implication in encoding the contingency value ΔP implies the knowledge of the conditional probabilities of the outcome given the action (Cheng, 1997; Hagmayer and Waldmann, 2007; Tanaka et al., 2008). Despite most of the literature implicates the OFC in the encoding of the stimulus-outcome associations, for example in response to the presentation of a cue signaling a reward (Salzman et al., 2007; Salzman and Fusi, 2010; Howard et al., 2015), we should consider that in instrumental learning, in order to establish the relation between the stimulus and its outcome, an agent should be able to link the information about actions in a stimulus-action-outcome association (O'Doherty, 2007). In a fMRI study (Valentin et al., 2007) conducted on human participants performing an outcome devaluation task, the results suggested that the OFC is able to represent actions-outcomes information, showing a different activation profile for valued and devalued actions. This result is also in line with animals' studies performed on rats showed that prefrontal cortex and dorso-medial striatum are important to learn actions-outcomes association during goal-directed learning (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Killcross, 2003). Moreover, the fact that OFC activity responds to ΔP and P(W|P) after outcome presentation, and responds to P(W|C) after the action executions, can highlight its role both in acquisition and update of the actions-outcomes association and in outcome prediction.

As the OFC, the prefrontal cortex (PFC) is implied in encoding outcome values. From our results we can see that the lateral rostro-ventral prefrontal cortex (PFrvl) seems to participate in encoding positive outcome values but only if associated to the action 'play' and not to any chosen action. The ventral prefrontal cortex is known to

mediate attentional processes and to encode stimulus salience (Asplund et al., 2010; Walther et al., 2011). Thus, we question whether this observed effect can be linked to an action dependent attentional mechanism, possibly derived by an unequal perception of the causal power attributable to the direct intervention of the agent (play) rather than a random environmental variable (no play). Further investigations about the role of this region in the attentional processes linked to instrumental learning are needed.

Concerning the P(W|P) and the P(W|C), we found significant activation also in the insular cortex (IC) after receiving the outcome. The IC is known to participate in instrumental behavior in encoding incentive memories together with the amygdala (Parkes et al., 2015) and in retrieving outcome incentive values in order to guide the actions, but not in learning action-outcomes associations (Parkes et al., 2017). Thus, the activation of this region responding to these two behavioral regressors after the outcome presentation can be linked to the update of these values.

Also the premotor rostro-ventral cortex (PMrv) seems to be involved in encoding the P(W|P). This result is particularly challenging to discuss, as we would expect to find a modulation of the PMrv before the action selection, participating in action planning (Gremel and Costa, 2013), and not after the outcome presentation. This area has been defined as a relay from parietal to medial prefrontal cortices in visuomotor task (Viejo et al., 2015), but also in this case further investigations are needed.

We observed an increment in MI in the parietal cortex in relation to ΔP values, however, statistical analysis showed that this increment is just below the significance threshold. Curiously, this effect turns out to be significant if we perform the analysis using the ΔP computed as log(P(W|P) / P(W|nP)) as behavioral regressors, while the effect found in the OFC is just below the significance threshold.

Together, these result seems to indicate a role of the parietal cortex in encoding contingency values, also according to previous literature showing parietal cortex tracks contingency values computed both as ΔP and as the Jensen-Shannon divergence between the probabilities of the outcome conditioned on different actions (Liljeholm et al., 2011, 2013).
Regarding the temporal lobe, its implication in instrumental learning is less understood in comparison to other regions, nonetheless its activity has been related to formation and updating inferences about optimal behavioral strategies (O'Doherty et al., 2017).

Despite most of the results presented here confirming previous findings, we believe that the neurophysiological correlates of goal-directed learning, and especially causal learning, needs a deeper investigation. In this work we focused specifically on the high-gamma band oscillatory activity, but it would be of particular interest to look into the contributions of other frequency bands, in order to better understand global computations associated to causal learning, as high-gamma activity reflects local computations, while lower frequency bands seems to be more associated to extensive computations (von Stein and Sarnthein, 2000).
The task that we proposed to participants is quite complex and requires more computational effort to be accomplished in comparison to a classical contingency learning task. One of the differences is that the participant is not called to choose between performing an action and not performing it, but rather on choosing one action or another. The taken decision is then transferred to a middle agent (the player under evaluation) that is then supposed to execute (or not) the action. Moreover the goal of the task is less explicit, as we ask the participants to maximize their knowledge about the performance of the player under evaluation, and not, for example, to maximise the number of achieved positive outcomes. This level of complexity offers us the opportunity to study several different aspects of causal

learning, but for these reasons, the significance of some element of this task should be better addressed both in modelling and in comparison with the brain activity.
Finally, It will be interesting to study the cortico-cortical interactions between pairs of brain regions forming functional networks supporting causal learning in time, using techniques such as Granger Causality (GC) or Partial Information Decomposition (PID).

6.4 Future perspectives

Besides the clinical purposes, the interest in neuroscience is to understand how the brain can successfully interact with the environment producing complex behaviors. These behaviors emerge from different levels of complexity, from the molecular interactions leading processes as the synaptic plasticity, to cellular organization, networks' dynamics and beyond. But this is not enough, indeed the environment can trigger changes inside the brain, both if we are interacting with it and if we are only observing it, activating networks, cells, and stimulating synaptic plasticity to form novel interactions. This is a two-way flux of information, a bottom-up process to produce successful behaviors, and a top-down one to learn. But lately, we also understood that the brain has not only a decoding/encoding function, but it is also generative, meaning that it is able to internally loop this information in order to produce predictions that can influence both behavior and learning. This is exactly what happens in goal-directed decision making: once setted a desired state, we observe the environment and we make a prediction that guides our response, then, depending on the feedback, an error signal is produced to update our prediction model. Less is still known on how we can represent goals, or internally generate goals; this would be a really interesting branch to explore both with models and with neurophysiological data analysis. For about seventy years to now, goal-directed learning and behavior have been studied at different levels, but what we still lack is to bind together this knowledge in order to fill the gaps. In my opinion, this can be

achieved by making some efforts in two directions: building biologically inspired models, and linking these models with brain activity.

Nowadays, powerful Bayesian and deep learning models are used to model brain behavior, and they work really well but despite that they're very poorly informative about the lower level computations. On the other side, we have the problem of computational power, indeed it is impossible to implement a neural network taking into account the whole complexity of the brain network that we are trying to model. Thus, it would be interesting to build models based on veritable neural architectures, able to perform higher bayesian computations as the state of the art models, finding a compromise between descriptive accuracy and computational performances.

Then, it would be interesting to use these models to explain neurophysiological recordings, similarly to what is done today on deep learning models, especially the one concerning vision and language processing, thanks to a recent technique called Representational Similarity Analysis (RSA) (Kriegeskorte, 2008).

Moreover, given the complexity on which this system relies, it would be of a certain interest to further investigate the interplay between cortical regions and basal ganglia and their functional connectivity and dynamics. This aim is harder to achieve because of the difficult accessibility of the subcortical regions, which allows a simultaneous accurate recording of cortical and subcortical regions only in animal models. More informations about cortico-striatal computations can also be useful to improve architectures and algorithms of GDB models

Bibliography

- Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. NeuroImage *31*, 790–795.
- Adams, C.D., and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. Q. J. Exp. Psychol. Sect. B *33*, 109–121.
- Alberquilla, S., Gonzalez-Granillo, A., Martín, E.D., and Moratalla, R. (2020). Dopamine regulates spine density in striatal projection neurons in a concentration-dependent manner. Neurobiol. Dis. *134*, 104666.
- Alkemade, A., and Forstmann, B.U. (2014). Do we need to revise the tripartite subdivision hypothesis of the human subthalamic nucleus (STN)? NeuroImage *95*, 326–329.
- Allan, L.G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. Bull. Psychon. Soc. *15*, 147–149.
- Allan, L.G. (1993). Human Contingency Judgments: Rule Based or Associative? Psychol. Bull. *114*, 435–448.
- Allan, L.G., and Jenkins, H.M. (1980). The judgment of contingency and the nature of the response alternatives. Can. J. Psychol. Can. Psychol. *34*, 1.
- Allan, L.G., Hannah, S.D., Crump, M.J.C., and Siegel, S. (2008). The psychophysics of contingency assessment. J. Exp. Psychol. Gen. *137*, 226–243.
- Apicella, P., Ljungberg, T., Scarnati, E., and Schultz, W. (1991). Responses to reward in monkey dorsal and ventral striatum. Exp. Brain Res. *85*.
- Ashby, F.G., Turner, B.O., and Horvitz, J.C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. Trends Cogn. Sci. *14*, 208–215.
- Asplund, C.L., Todd, J.J., Snyder, A.P., and Marois, R. (2010). A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. Nat. Neurosci. *13*, 507–512.
- Assous, M., and Tepper, J.M. (2019). Excitatory extrinsic afferents to striatal interneurons and interactions with striatal microcircuitry. Eur. J. Neurosci. *49*, 593–603.
- Auzias, G., Coulon, O., and Brovelli, A. (2016). MarsAtlas: A cortical parcellation atlas for functional mapping. Hum. Brain Mapp. *37*, 1573–1592.
- Baker, C., Saxe, R., and Tenenbaum, J.B. (2006). Bayesian models of human action understanding. 8.
- Baldassarre, G., Lord, W., Granato, G., and Santucci, V.G. (2019). An Embodied Agent Learning Affordances With Intrinsic Motivations and Solving Extrinsic Tasks With Attention and One-Step Planning. Front. Neurorobotics *13*.
- Balleine, B.W., and Dickinson, A. (1998). Goal-directed instrumental action:

contingency and incentive learning and their cortical substrates. Neuropharmacology *37*, 407–419.
- Balleine, B.W., and O'Doherty, J.P. (2010). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacology *35*, 48–69.
- Balleine, B.W., Delgado, M.R., and Hikosaka, O. (2007). The Role of the Dorsal Striatum in Reward and Decision-Making. J. Neurosci. *27*, 8161–8165.
- Balleine, B.W., Dezfouli, A., Ito, M., and Doya, K. (2015). Hierarchical control of goal-directed action in the cortical–basal ganglia network. Curr. Opin. Behav. Sci. *5*, 1–7.
- Basanisi, R., Brovelli, A., Cartoni, E., and Baldassarre, G. (2020). A generative spiking neural-network model of goal-directed behaviour and one-step planning. PLOS Comput. Biol. *16*, e1007579.
- Bird, C.M., and Burgess, N. (2008). The hippocampus and memory: insights from spatial processing. Nat. Rev. Neurosci. *9*, 182–194.
- Bissonette, G.B., and Roesch, M.R. (2015). Rule encoding in dorsal striatum impacts action selection. Eur. J. Neurosci. *42*, 2555–2567.
- Blaisdell, A.P. (2006). Causal Reasoning in Rats. Science *311*, 1020–1022.
- Bonet, B., and Geffner, H. (2014). Planning with Incomplete Information as Heuristic Search in Belief Space. 10.
- Botvinick, M.M., Niv, Y., and Barto, A.G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. Cognition *113*, 262–280.
- Brovelli, A., Lachaux, J.-P., Kahane, P., and Boussaoud, D. (2005). High gamma frequency oscillatory activity dissociates attention from intention in the human premotor cortex. NeuroImage *28*, 154–164.
- Brovelli, A., Laksiri, N., Nazarian, B., Meunier, M., and Boussaoud, D. (2008). Understanding the Neural Computations of Arbitrary Visuomotor Learning through fMRI and Associative Learning Theory. Cereb. Cortex *18*, 1485–1495.
- Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons. PLoS Comput. Biol. *7*, e1002211.
- Buzsáki, G. (2006). Rhythms of the Brain (Oxford University Press).
- Buzsáki, G., Anastassiou, C.A., and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. *13*, 407–420.
- Cheng, P.W. (1997). From covariation to causation: a causal power theory. Psychol. Rev. *104*, 367.
- Chib, V.S., Rangel, A., Shimojo, S., and O'Doherty, J.P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. J. Neurosci. *29*, 12315–12320.
- Cole, S.R., and Voytek, B. (2017). Brain Oscillations and the Importance of Waveform Shape. Trends Cogn. Sci. *21*, 137–149.

- Corbetta, M., and Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. *3*, 201–215.
- Corbit, L.H., and Balleine, B.W. (2003). The role of prelimbic cortex in instrumental conditioning. Behav. Brain Res. *146*, 145–157.
- Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.
- Dayan, P., Hinton, G.E., Neal, R.M., and Zemel, R.S. (1995). The Helmholtz Machine. Neural Comput. *7*, 889–904.
- De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. Learn. Behav. *37*, 1–20.
- Delgado, M.R., Miller, M.M., Inati, S., and Phelps, E.A. (2005). An fMRI study of reward-related probability learning. NeuroImage *24*, 862–873.
- Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage *31*, 968–980.
- Dezfouli, A., and Balleine, B.W. (2013). Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. PLoS Comput. Biol. *9*, e1003364.
- Dickinson, A. (1985). Actions and Habits: The Development of Behavioural Autonomy. Philos. Trans. R. Soc. B Biol. Sci. *308*, 67–78.
- Dickinson, A. (2012). Associative learning and animal cognition. Philos. Trans. R. Soc. B Biol. Sci. *367*, 2733–2742.
- Dickinson, A., and Balleine, B. (1994). Motivational control of goal-directed action. Anim. Learn. Behav. *22*, 1–18.
- Dickinson, A., and Balleine, B.W. (2000). Causal Cognition and Goal-Directed Action. Evol. Cogn. 185.
- Dolan, R.J., and Dayan, P. (2013). Goals and Habits in the Brain. Neuron *80*, 312–325.
- Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., and Daw, N.D. (2015). Model-based choices involve prospective neural activity. Nat. Neurosci. *18*, 767–772.
- Domenech, P., Rheims, S., and Koechlin, E. (2020). Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. Science *369*, eabb0184.
- Dotson, N.M., Hoffman, S.J., Goodell, B., and Gray, C.M. (2017). A Large-Scale Semi-Chronic Microdrive Recording System for Non-Human Primates. Neuron *96*, 769-782.e2.
- Eisinger, R.S., Urdaneta, M.E., Foote, K.D., Okun, M.S., and Gunduz, A. (2018). Non-motor Characterization of the Basal Ganglia: Evidence From Human and Non-human Primate Electrophysiology. Front. Neurosci. *12*.
- Eisinger, R.S., Cernera, S., Gittis, A., Gunduz, A., and Okun, M.S. (2019). A

review of basal ganglia circuits and physiology: Application to deep brain stimulation. Parkinsonism Relat. Disord. *59*, 9–20.
- Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., and Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. Front. Artif. Intell. *3*.
- Engel, A.K., and Fries, P. (2010). Beta-band oscillations—signalling the status quo? Curr. Opin. Neurobiol. *20*, 156–165.
- Feingold, J., Gibson, D.J., DePasquale, B., and Graybiel, A.M. (2015). Bursts of beta oscillation differentiate postperformance activity in the striatum and motor cortex of monkeys performing movement tasks. Proc. Natl. Acad. Sci. *112*, 13687–13692.
- Freedman, D.J., and Assad, J.A. (2006). Experience-dependent representation of visual categories in parietal cortex. Nature *443*, 85–88.
- Friedrich, J., and Lengyel, M. (2016). Goal-Directed Decision Making with Spiking Neurons. J. Neurosci. *36*, 1529–1546.
- Ghosh-Dastidar, S., and Adeli, H. (2009). Spiking Neural Networks. 14.
- Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron *66*, 585–595.
- Gollwitzer, P.M., and Moskowitz, G.B. Goal Effects on Action and Cognition. 39.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. 9.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning.
- Graybiel, A.M. (2008). Habits, Rituals, and the Evaluative Brain. Annu. Rev. Neurosci. *31*, 359–387.
- Gremel, C.M., and Costa, R.M. (2013). Premotor cortex is critical for goal-directed actions. Front. Comput. Neurosci. *7*.
- Griffiths, T.L., Kemp, C., and Tenenbaum, J.B. (2008). Bayesian models of cognition.
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. Pattern Recognit. *77*, 354–377.
- Gueguen, M.C.M., Lopez-Persem, A., Billeke, P., Lachaux, J.-P., Rheims, S., Kahane, P., Minotti, L., David, O., Pessiglione, M., and Bastin, J. (2021). Anatomical dissociation of intracerebral signals for reward and punishment prediction errors in humans. Nat. Commun. *12*.
- Haber, S.N. (2003). The primate basal ganglia: parallel and integrative networks. J. Chem. Neuroanat. *26*, 317–330.
- Hagmayer, Y., and Waldmann, M.R. (2007). Inferences about unobserved causes in human contingency learning. Q. J. Exp. Psychol. *60*, 330–355.
- Hammond, L.J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. J. Exp. Anal. Behav. *34*, 297–304.

- Han, M.-J., Park, C.-U., Kang, S., Kim, B., Nikolaidis, A., Milham, M.P., Hong, S.J., Kim, S.-G., and Baeg, E. (2021). Mapping functional gradients of the striatal circuit using simultaneous microelectric stimulation and ultrahigh-field fMRI in non-human primates. NeuroImage *236*, 118077.
- Hilario, M.R.F. (2008). High on habits. Front. Neurosci. *2*, 208–217.
- Holmes, N.M., Marchand, A.R., and Coutureau, E. (2010). Pavlovian to instrumental transfer: A neurobehavioural perspective. Neurosci. Biobehav. Rev. *34*, 1277–1295.
- Holt, A.B., Kormann, E., Gulberti, A., Pötter-Nerger, M., McNamara, C.G., Cagnan, H., Baaske, M.K., Little, S., Köppen, J.A., Buhmann, C., et al. (2019). Phase-Dependent Suppression of Beta Oscillations in Parkinson's Disease Patients. J. Neurosci. *39*, 1119–1134.
- Howard, J.D., Gottfried, J.A., Tobler, P.N., and Kahnt, T. (2015). Identity-specific coding of future rewards in the human orbitofrontal cortex. Proc. Natl. Acad. Sci. *112*, 5195–5200.
- Hultborn, H. (2006). Spinal reflexes, mechanisms and concepts: From Eccles to Lundberg and beyond. Prog. Neurobiol. *78*, 215–232.
- Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., and Schyns, P.G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula: Gaussian Copula Mutual Information. Hum. Brain Mapp. *38*, 1541–1573.
- Izawa, J., and Shadmehr, R. (2011). Learning from Sensory and Reward Prediction Errors during Motor Adaptation. PLoS Comput. Biol. *7*, e1002012.
- Jahanshahi, M., Obeso, I., Rothwell, J.C., and Obeso, J.A. (2015). A fronto–striato–subthalamic–pallidal network for goal-directed and habitual inhibition. Nat. Rev. Neurosci. *16*, 719–732.
- Jenkinson, N., and Brown, P. (2011). New insights into the relationship between dopamine, beta oscillations and motor function. Trends Neurosci. *34*, 611–618.
- Kappel, D., Nessler, B., and Maass, W. (2014). STDP Installs in Winner-Take-All Circuits an Online Approximation to Hidden Markov Model Learning. PLoS Comput. Biol. *10*, e1003511.
- Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. PLoS Comput. Biol. *7*, e1002055.
- Khanna, P., and Carmena, J.M. (2017). Beta band oscillations in motor cortex reflect neural population signals that delay movement onset. ELife *6*.
- Killcross, S. (2003). Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. Cereb. Cortex *13*, 400–408.
- Kingma, D.P., and Welling, M. (2014). Auto-Encoding Variational Bayes. ArXiv13126114 Cs Stat.
- Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci.

- Kühn, A.A., Williams, D., Kupsch, A., Limousin, P., Hariz, M., Schneider, G., Yarrow, K., and Brown, P. (2004). Event-related beta desynchronization in human subthalamic nucleus correlates with motor performance. Brain *127*, 735–746.
- Lanciego, J.L., Luquin, N., and Obeso, J.A. (2012). Functional Neuroanatomy of the Basal Ganglia. Cold Spring Harb. Perspect. Med. *2*, a009621–a009621.
- Lecumberri, A., Lopez-Janeiro, A., Corral-Domenge, C., and Bernacer, J. (2017). Neuronal density and proportion of interneurons in the associative, sensorimotor and limbic human striatum. Brain Struct. Funct.
- LeDoux, J., Iwata, J., Cicchetti, P., and Reis, D. (1988). Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. J. Neurosci. *8*, 2517–2529.
- Lee, S.W., Shimojo, S., and O'Doherty, J.P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. Neuron *81*, 687–699.
- Leventhal, D.K., Gage, G.J., Schmidt, R., Pettibone, J.R., Case, A.C., and Berke, J.D. (2012). Basal Ganglia Beta Oscillations Accompany Cue Utilization. Neuron *73*, 523–536.
- Levy, D.J., and Glimcher, P.W. (2012). The root of all value: a neural common currency for choice. Curr. Opin. Neurobiol. *22*, 1027–1038.
- Levy, R., and Goldman-Rakic, P.S. (2000). Segregation of working memory functions within the dorsolateral prefrontal cortex. Exp. Brain Res. *133*, 23–32.
- Li, L., Du, Y., Li, N., Wu, X., and Wu, Y. (2009). Top–down modulation of prepulse inhibition of the startle reflex in humans and rats. Neurosci. Biobehav. Rev. *33*, 1157–1167.
- Liljeholm, M. (2018). Instrumental Divergence and Goal-Directed Choice. In Goal-Directed Decision Making, (Elsevier), pp. 27–48.
- Liljeholm, M. (2021). Agency and goal-directed choice. Curr. Opin. Behav. Sci. *41*, 78–84.
- Liljeholm, M., and O'Doherty, J.P. (2012). Contributions of the striatum to learning, motivation, and performance: an associative account. Trends Cogn. Sci. *16*, 467–475.
- Liljeholm, M., Tricomi, E., O'Doherty, J.P., and Balleine, B.W. (2011). Neural Correlates of Instrumental Contingency Learning: Differential Effects of Action-Reward Conjunction and Disjunction. J. Neurosci. *31*, 2474–2480.
- Liljeholm, M., Wang, S., Zhang, J., and O'Doherty, J.P. (2013). Neural Correlates of the Divergence of Instrumental Probability Distributions. J. Neurosci. *33*, 12519–12527.
- Lipton, Z.C., Berkowitz, J., and Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. ArXiv150600019 Cs.
- Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. Neural Netw. *10*, 1659–1671.

- Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal Correlates of Goal-Based Motor Selection in the Prefrontal Cortex. Science *301*, 229–232.
- McDannald, M.A., Lucantonio, F., Burke, K.A., Niv, Y., and Schoenbaum, G. (2011). Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning. J. Neurosci. *31*, 2700–2705.
- van der Meer, M. (2010). Integrating early results on ventral striatal gamma oscillations in the rat. Front. Neurosci.
- Mehlhorn, K., Newell, B.R., Todd, P.M., Lee, M.D., Morgan, K., Braithwaite, V.A., Hausmann, D., Fiedler, K., and Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision *2*, 191–215.
- Mestres-Missé, A., Turner, R., and Friederici, A.D. (2012). An anterior–posterior gradient of cognitive control within the dorsomedial striatum. NeuroImage *62*, 41–47.
- Miller, E.K., and Cohen, J.D. (2001). An Integrative Theory of Prefrontal Cortex Function. Annu. Rev. Neurosci. *24*, 167–202.
- Mitchell, C.J., De Houwer, J., and Lovibond, P.F. (2009). The propositional nature of human associative learning. Behav. Brain Sci. *32*, 183.
- Morris, L.S., Kundu, P., Dowell, N., Mechelmans, D.J., Favre, P., Irvine, M.A., Robbins, T.W., Daw, N., Bullmore, E.T., Harrison, N.A., et al. (2016). Fronto-striatal organization: Defining functional and microstructural substrates of behavioural flexibility. Cortex *74*, 118–133.
- Morris, R.W., Dezfouli, A., Griffiths, K.R., Le Pelley, M.E., and Balleine, B.W. (2017). The algorithmic neuroanatomy of action-outcome learning (Neuroscience).
- Moser, E.I., Kropff, E., and Moser, M.-B. (2008). Place Cells, Grid Cells, and the Brain's Spatial Representation System. Annu. Rev. Neurosci. *31*, 69–89.
- Münte, T.F., Marco-Pallares, J., Bolat, S., Heldmann, M., Lütjens, G., Nager, W., Müller-Vahl, K., and Krauss, J.K. (2017). The human globus pallidus internus is sensitive to rewards – Evidence from intracerebral recordings. Brain Stimulat. *10*, 657–663.
- Nakano, K., Kayahara, T., Tsutsumi, T., and Ushiro, H. (2000). Neural circuits and functional organization of the striatum. J. Neurol. *247*, V1–V15.
- Newman, M.E.J. (2011). Complex Systems: A Survey. Am. J. Phys. *79*, 800–810.
- O'Doherty, J.P. (2007). Lights, Camembert, Action! The Role of Human Orbitofrontal Cortex in Encoding Stimuli, Rewards, and Choices. Ann. N. Y. Acad. Sci. *1121*, 254–272.
- O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., and Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. Nat. Neurosci. *4*, 95–102.
- O'Doherty, J.P., Hampton, A., and Kim, H. (2007). Model-Based fMRI and Its

Application to Reward Learning and Decision Making. Ann. N. Y. Acad. Sci. *1104*, 35–53.
- O'Doherty, J.P., Cockburn, J., and Pauli, W.M. (2017). Learning, reward, and decision making. Annu. Rev. Psychol. *68*, 73–100.
- Ostlund, S.B., Winterbauer, N.E., and Balleine, B.W. (2009). Evidence of Action Sequence Chunking in Goal-Directed Instrumental Conditioning and Its Dependence on the Dorsomedial Prefrontal Cortex. J. Neurosci. *29*, 8280–8287.
- Oya, H., Adolphs, R., Kawasaki, H., Bechara, A., Damasio, A., and Howard, M.A. (2005). Electrophysiological correlates of reward prediction error recorded in the human prefrontal cortex. Proc. Natl. Acad. Sci. *102*, 8351–8356.
- Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature *441*, 223–226.
- Palminteri, S., and Pessiglione, M. (2017). Opponent Brain Systems for Reward and Punishment Learning. In Decision Neuroscience, (Elsevier), pp. 291–303.
- Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., et al. (2012). Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. Neuron *76*, 998–1009.
- Parkes, S.L., Bradfield, L.A., and Balleine, B.W. (2015). Interaction of Insular Cortex and Ventral Striatum Mediates the Effect of Incentive Memory on Choice Between Goal-Directed Actions. J. Neurosci. *35*, 6464–6471.
- Parkes, S.L., Ravassard, P.M., Cerpa, J.-C., Wolff, M., Ferreira, G., and Coutureau, E. (2017). Insular and Ventrolateral Orbitofrontal Cortices Differentially Contribute to Goal-Directed Behavior in Rodents. Cereb. Cortex 1–13.
- Paton, J.J., Belova, M.A., Morrison, S.E., and Salzman, C.D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. Nature *439*, 865–870.
- Pavlov, I.P., and Anrep, G.V. (1927). Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex.
- Penn, D.C., and Povinelli, D.J. (2007). Causal Cognition in Human and Nonhuman Animals: A Comparative, Critical Review. Annu. Rev. Psychol. *58*, 97–118.
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042–1045.
- Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. Nature *497*, 74–79.
- Pizzo, F., Roehri, N., Medina Villalon, S., Trébuchon, A., Chen, S., Lagarde, S., Carron, R., Gavaret, M., Giusiano, B., McGonigal, A., et al. (2019). Deep brain

activities can be detected with magnetoencephalography. Nat. Commun. *10*.
- Plassmann, H., O'Doherty, J.P., and Rangel, A. (2010). Appetitive and Aversive Goal Values Are Encoded in the Medial Orbitofrontal Cortex at the Time of Decision Making. J. Neurosci. *30*, 10799–10808.
- Platt, M.L., and Glimcher, P.W. (1999). Neural correlates of decision variables in parietal cortex. Nature *400*, 233–238.
- Ponulak, F., and Kasinski, A. (2011). Introduction to spiking neural networks: Information processing, learning and applications. Acta Neurobiol. Exp. (Warsz.) *4*.
- Procyk, E., and Goldman-Rakic, P.S. (2006). Modulation of Dorsolateral Prefrontal Delay Activity during Self-Organized Behavior. J. Neurosci. *26*, 11313–11323.
- Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M.C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M.R., and Obeso, J.A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. Nat. Rev. Neurosci. *11*, 760–772.
- Rescorla, R.A. (1966). Predictability and number of pairings in Pavlovian fear conditioning. Psychon. Sci. *4*, 383–384.
- Rescorla, R.A. (1968). Probability of shock in the presence and absence of cs in fear conditioning. J. Comp. Physiol. Psychol. *66*, 1–5.
- Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement (Meredith Corporation).
- Ressler, N. (2004). Rewards and punishments, goal-directed behavior and consciousness. Neurosci. Biobehav. Rev. *28*, 27–39.
- Rizley, R.C., and Rescorla, R.A. (1972). Associations in second-order conditioning and sensory preconditioning. J. Comp. Physiol. Psychol. *81*, 1–11.
- Rolls, E.T. (2003). Representations of Pleasant and Painful Touch in the Human Orbitofrontal and Cingulate Cortices. Cereb. Cortex *13*, 308–317.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev. *65*, 386–408.
- Rueckert, E., Kappel, D., Tanneberg, D., Pecevski, D., and Peters, J. (2016). Recurrent Spiking Networks Solve Planning Tasks. Sci. Rep. *6*.
- Salzman, C.D., and Fusi, S. (2010). Emotion, Cognition, and Mental State Representation in Amygdala and Prefrontal Cortex. Annu. Rev. Neurosci. *33*, 173–202.
- Salzman, C.D., Paton, J.J., Belova, M.A., and Morrison, S.E. (2007). Flexible Neural Representations of Value in the Primate Brain. Ann. N. Y. Acad. Sci. *1121*, 336–354.
- Schoenbaum, G., Chiba, A.A., and Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. Nat. Neurosci. *1*, 155–159.

- Schreiner, D.C., Renteria, R., and Gremel, C.M. (2020). Fractionating the all-or-nothing definition of goal-directed and habitual decision-making. J. Neurosci. Res. *98*, 998–1006.
- Schultz, W. (2007). Multiple Dopamine Functions at Different Time Courses. Annu. Rev. Neurosci. *30*, 259–288.
- Schultz, W. (2016a). Dopamine reward prediction error coding. Dialogues Clin. Neurosci. *18*, 10.
- Schultz, W. (2016b). Dopamine reward prediction-error signalling: a two-component response. Nat. Rev. Neurosci. *17*, 183–195.
- Schultz, W., Dayan, P., and Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. *275*, 30.
- Schwerdt, H.N., Amemori, K., Gibson, D.J., Stanwicks, L.L., Yoshida, T., Bichot, N.P., Amemori, S., Desimone, R., Langer, R., Cima, M.J., et al. (2020). Dopamine and beta-band oscillations differentially link to striatal value and motor control. Sci. Adv. *6*, eabb9226.
- Seeber, M., Cantonas, L.-M., Hoevels, M., Sesia, T., Visser-Vandewalle, V., and Michel, C.M. (2019). Subcortical electrophysiological activity is detectable with high-density EEG source imaging. Nat. Commun. *10*.
- Shadlen, M.N., and Newsome, W.T. (2001). Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. J. Neurophysiol. *86*, 1916–1936.
- Shanks, D.R. (1995). The Psychology of Associative Learning (Cambridge University Press).
- Shannon, C.E. (1948). A Mathematical Theory of Communication. *27*, 379–423.
- Shrestha, A., and Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures. IEEE Access *7*, 53040–53065.
- Skinner, B.F. (1938). The Behavior of Organisms (Oxford, England: Appleton-Century).
- Smith, D.V., Hayden, B.Y., Truong, T.-K., Song, A.W., Platt, M.L., and Huettel, S.A. (2010). Distinct Value Signals in Anterior and Posterior Ventromedial Prefrontal Cortex. J. Neurosci. *30*, 2490–2495.
- von Stein, A., and Sarnthein, J. (2000). Different frequencies for different scales of cortical integration: from local gamma to long range alphartheta synchronization. Int J Psychophysiol *38*, 301–313.
- Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: an introduction.
- Takikawa, Y., Kawagoe, R., Itoh, H., Nakahara, H., and Hikosaka, O. (2002). Modulation of saccadic eye movements by predicted reward outcome. Exp. Brain Res. *142*, 284–291.
- Tanaka, S.C., Balleine, B.W., and O'Doherty, J.P. (2008). Calculating Consequences: Brain Systems That Encode the Causal Effects of Actions. J. Neurosci. *28*, 6750–6755.
- Tanneberg, D., Paraschos, A., Peters, J., and Rueckert, E. (2016). Deep spiking

networks for model-based planning in humanoids. In Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference On, (IEEE), pp. 656–661.
- Thorndike, E.L. (1898). Animal intelligence: An experimental study of the associative processes in animals. Psychol. Rev. Monogr. Suppl. *2*, i.
- Timme, N.M., and Lapish, C. (2018). A Tutorial for Information Theory in Neuroscience. Eneuro *5*, ENEURO.0052-18.2018.
- Tolman, E.C. (1948). Cognitive maps in rats and men. *55*, 189–208.
- Tschantz, A., Seth, A.K., and Buckley, C.L. (2020). Learning action-oriented models through active inference. PLOS Comput. Biol. *16*, e1007805.
- Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the Neural Substrates of Goal-Directed Learning in the Human Brain. J. Neurosci. *27*, 4019–4026.
- Viejo, G., Khamassi, M., Brovelli, A., and Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. Front. Behav. Neurosci. *9*.
- Vogelsang, D.A., and D'Esposito, M. (2018). Is There Evidence for a Rostral-Caudal Gradient in Fronto-Striatal Loops and What Role Does Dopamine Play? Front. Neurosci. *12*.
- Walther, S., Friederich, H.-C., Stippich, C., Weisbrod, M., and Kaiser, S. (2011). Response inhibition or salience detection in the right ventrolateral prefrontal cortex? NeuroReport *22*, 778–782.
- Wang, K.S., Smith, D.V., and Delgado, M.R. (2016). Using fMRI to study reward processing in humans: past, present, and future. J. Neurophysiol. *115*, 1664–1678.
- Wasserman, E.A., and Miller, R.R. (1997). What's elementary about associative learning? Annu. Rev. Psychol. *48*, 573–607.
- Watkins, C.J., and Dayan, P. (1992). Q-learning. Mach. Learn. *8*, 279–292.
- Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal Cortex as a Cognitive Map of Task Space. Neuron *81*, 267–279.
- Wimmer, G.E., and Shohamy, D. (2012). Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. Science *338*, 270–273.
- Yager, L.M., Garcia, A.F., Wunsch, A.M., and Ferguson, S.M. (2015). The ins and outs of the striatum: Role in drug addiction. Neuroscience *301*, 529–541.
- Yin, H.H., and Knowlton, B.J. (2006). The role of the basal ganglia in habit formation. Nat. Rev. Neurosci. *7*, 464–476.
- (2007). Bayesian brain: probabilistic approaches to neural coding (Cambridge, Mass: MIT Press).