ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE
SCHOOL OF LIFE SCIENCES

EPFL

Master's thesis in Life Sciences Engineering

Towards the neural basis of belief state computation in the dopamine system

Conducted by

Célia Benquet

Under the supervision of Prof. Naoshige Uchida,
Head of the Uchida Laboratory,
Harvard University, Cambridge, US

HARVARD
UNIVERSITY

Under the direction of Prof. Mackenzie Mathis,
Head of the Mathis Lab of Adaptive Motor Control,
EPFL, Lausanne, Switzerland

with the support of

EPFL • WISH
FOUNDATION
WOMEN IN SCIENCE AND HUMANITIES

LAUSANNE, EPFL, SPRING 2022

Acknowledgements

You're the average of the people who surround you. I've found it to be particularly true during my time at university, considering the amount of passionate and inspiring people I met along the way. I am eternally grateful for all of them. They have made my average grow so much and are responsible for the person I am now.

I would like to acknowledge Sandra and Jay without whom this project wouldn't have been possible. Thank you for your scientific guidance, interest and our numerous conversations. I also thank Nao for taking a chance on me. Thank you for giving me the opportunity to spend some time in your lab and to discover the academic world like I never though I would have. Thanks also to the Uchida lab members for welcoming me so warmly those past 6 months and for being such amazing coworkers and - I would even like to think - friends.

I also would like to express my gratitude to Mackenzie. I couldn't have dream of a better way to be introduced to the academic world. You're an inspiring person, and even more as a young woman scientist, I'm really glad I got to meet you.

Thank you to the EPFL WISH Foundation for their financial support, allowing me to write my Master's thesis abroad, in such a prestigious institution. I thank them for believing in me and my project and more generally in scientific women. I truly believe it is with foundations like this one that we can tackle the place of women in society and change the norm. Education is key!

Last but - definitely! - not least, thank you to my amazing family and friends for being there with me along the - sometimes tedious, but most of the time incredibly fun - way. Special shoot-out to my parents, for their ever-lasting support and amazing interest in my studies. You, more than anyone else, know how much coming to Boston during the course of my studies was a dream come true so thank you for making it possible.

Contents

List of Figures

List of Tables

Acronyms

ACG Autocorrelogram.
ANN artificial neural network.
AUROC area under the receiver operating characteristic curve.

BOLD blood oxygen level-dependent.

CCF Common Coordinate Frameworks.
CCG Crosscorrelogram.
CS conditioned stimulus.
CSC complete serial compound.

DA dopamine.
DP dynamic programming.

ELBO evidence lower bound.

GLM General Linear Model.
GRU gated recurrent unit.

HMM hidden markov model.

ISI inter-stimulus interval.
ITI inter-trial interval.

KORD κ-opioid receptors (KOR)-DREADD.

LDS linear dynamical systems.
LSTM long short-term memory.

M1 primary motor cortex.
M2 secondary motor cortex.
MDP Markov decision process.
MO motor areas.
mPFC medial prefrontal cortex.
MSE mean squared error.

OFC orbito-frontal cortex.

PCA principal component analysis.
PFC pre-frontal cortex.
PIR piriform.
POMDP partially-observable Markov decision process.
PSTH peristimulus time histogram.

R-W Rescorla-Wagner.
RL reinforcement learning.
RNN recurrent neural network.
RPE reward prediction error.
rSLDS recurrent switching linear dynamical systems.

SalB salvinorin B.
SE state estimator.
SLDS switching linear dynamical systems.
SN substantia nigra.
SNc substantia nigra pars compacta.
SVD singular value decomposition.
SVM support vector machines.

TD temporal difference.

US unconditioned stimulus.

VTA ventral tegmental area.

Abstract

Naturalistic decision-making is most of the time uncertain. Animals must learn to select actions based on noisy sensory information and incomplete knowledge of the world. To that extent, it has been proposed that the brain infers a probability distribution, or *belief state*, given the history of observations and actions, over hidden states of the environment. In fact, in partially-observable environments, dopamine (DA) response in the brain is better explained by the traditional temporal difference (TD) learning model implicating beliefs rather than fully observable states. In addition, the orbito-frontal cortex (OFC) was found to play a critical role in uncertain decision-making and is thought to encode belief state representation in the brain. However, whether, and, if such, how belief states are represented is not clear.

In this project, we use the variable reward delay task to tackle the neural basis of belief state computation in the OFC during reward learning with state uncertainty. In this classical conditioning paradigm, the reward is delivered either in 100% (task 1) or 90% (task 2) of the trials. The mice learn to associate cues to specific reward delays, if reward happens. We recorded neurons in the OFC for task 1 and interpreted our results at the single-neuron and population levels. Neural activity of individual neurons showed a large diversity of patterns. Each neuron had a specific pattern of activity and was tuned to relevant events of the task or to a specific time step. Hence, neurons taken as a population (1) were able to keep track of time elapsed and (2) showed two attractor-like fixed-points driving the dynamics of the neural activity, respectively when waiting for reward or in between trials.

We propose that the belief state is encoded in the OFC at the population level, by integrating explicit signals of the individual neurons. Each is tuned to specific state-relevant elements of the environment. Further work on task 2 should be performed to assess our proposal.

Keywords: Decision-making, Uncertainty, OFC, Belief state representation, Attractor-like dynamics

Chapter 1

Introduction

New England's weather is unpredictable! One day the sun shines and birds are singing, the next day it is pouring rain or, even worst, snowing. To decide on what to wear, relying only on an observation of the current weather to predict the weather for the rest of the day is not sufficient. Dressing to go to work becomes an interesting guessing game, trying to balance between the weather and temperature forecasts, the current observation on the weather through the window and one's unpleasant experience on the previous evening, getting caught in the rain unexpectedly. On top of that, someone who experienced going home soaked the previous day might be more precautionary than someone who avoided the rain, by going home later for instance. With experience, one develops their posterior estimates of the coming weather, i.e. $P(rainy\_day|observations)$ and $P(sunny\_day|observations) = 1 - P(rainy\_day|observations)$, considering their own prior experiences, which might differ from one person to the other.

A similar process is thought to be at the origin of most of the common, simple forms of decision-making in the real world. Animals learn to predict and control future rewards based on past experiences and current, often unreliable and incomplete, sensory stimuli by associating that information to the outcome value. In that sense, reinforcement learning (RL) theory provides a powerful computational framework to understand how the brain solves decision problems [1]. In particular, human and animal experiments strongly support the hypothesis that the phasic activity of midbrain dopamine (DA) neurons encodes a reward prediction error (RPE), strikingly similarly to the errors of the TD learning model [2][3][4]. In that model, RPE - the difference between received and predicted reward - is used to update predictions and improve their accuracy [5].

However, most RL theories of the DA system assume that the current state of the environment is fully observable, while this assumption is untenable in the real world. Computational theories propose that dealing with state uncertainty requires the computation of a so-called *belief state*: a probability distribution inferred over all possible states and estimating the current hidden state [6]. Building on this, previous work has suggested that, when states are uncertain, the RPEs signaled by DA neurons are calculated with respect to belief state [7][8][9]. Direct empirical evidence for this theory has been sparse [10][11] and some of the same data can be explained by alternative frameworks (e.g., Ludvig et al., 2008 [12]; Gershman et al., 2014 [13]). However, two experimental paradigms recently developed by the Uchida/Gershman labs present evidence that the brain computes RPE using a belief state [14][15]. Furthermore, medial prefrontal cortex (mPFC) inactivation in Pavlovian conditioning paradigm involving state uncertainty suggested that this region plays a critical role in either or both computing and representing belief state [16]. The precise nature of this contribution, however, remains unclear.

Therefore, there are several outstanding questions as to *where* belief states are represented in the brain and *how* neural circuits can represent them.

In this thesis, we will investigate belief state encoding in the mouse OFC, a region known to be implicated in state representation. More particularly, we will study the dynamics of those belief states during a reward learning task with state uncertainty. For that, we will use neuronal activity recorded in mice during a variable reward delay task [14].

First, we will analyze single-neurons activity in the OFC and compare it to different regions of the prefrontal cortex, namely the primary motor cortex (M1), secondary motor cortex (M2) and piriform (PIR) cortex. For that, we will observe individual neurons as they display typical patterns of activity. Then, we will examine whether external and behavioral variables from the task are encoded in neuronal activity, using a linear regression model. Variables that are most likely encoded by the neurons are those that are useful to predict neuronal firing rates in that model. Finally, we will assess whether neuronal activity resemble beliefs. For that, we will train a linear classifier to predict the most probable state at a given time step from the neural firing rate and compare performances to the true belief state.

In the second part, we will investigate the dynamics of neural population activity. A preliminary study showed that the dynamics of a gated recurrent unit (GRU) network trained on the task to predict value displayed two fixed-points attractors corresponding to the states of the environment. We will show the resemblance between the dynamics of such a network and the neural population activity in the OFC, using an unsupervised dynamical systems framework [17].

1.1 Basis of reinforcement learning (RL) and the Markov decision process (MDP)

Sutton and Barto define RL as a way to learn *what to do* to optimally perform a task by maximizing *reward*, through *trial-and-error* [5]. A decision-making agent interacts with its environment by taking actions in it to achieve a goal. Then, it can use its experiences from past trials to selecting the optimal action on the current trial and improve performances - i.e. increase reward. This way, it learns - and improves constantly - a *policy*, that defines how the agent will behave in a given *state* of the world. A *state* consists of information about the environment at a given time, such that it is *sufficient* and *relevant* to predict rewards and transitions between states in an optimal way [18]. It can consist of the place, objects, or time since some previous events, for instance, but it is not a one-on-one reflection of the physical state of the environment. The type of sufficient and relevant information will vary depending on the task to achieve. The Markov decision process (MDP) framework (Fig. 1.1) formally defines those interactions in terms of states, actions, transition probabilities and reward (S, A, T, R), where:

- S is a finite *state space*;
- A is a finite *action space* with A_s the set of actions available from state s;
- T is a *state-transition function* so that $T_a(s, s') = Pr(s_{t+1} = s'|s_t = s, a_t = a)$ is the probability that action a taken in state s at time t leads to state s' at time $t + 1$.
- R is the *reward function* so that $R_a(s, s')$ is the intermediate reward received after transitioning from s to s', taking action a.

States and rewards are assumed to follow a Markov process, such that state transitions and rewards are conditionally independent of past events given the current state [5]. The value function $V(s)$ can consequently be expressed as a recursive expression, known as *Bellman's equation* [19]:

$$V(s) = R(s) + \gamma E[V(s')] \tag{1.1}$$

with $R(s)$ the immediate reward obtained upon visiting state s, γ, a discount factor that down-weights future rewards exponentially ($0 \leq \gamma \leq 1$), and $E[V(s')]$ the expected value at next state s', which recursively includes expectations of future rewards in its calculation.

Once the context is defined as an MDP, diverse algorithms can be applied to learn the task. One of them has had a lot of attention in neurosciences: the TD learning model. It is a class of model-free methods learning both from experience and bootstrapping [20]. The current value is not only

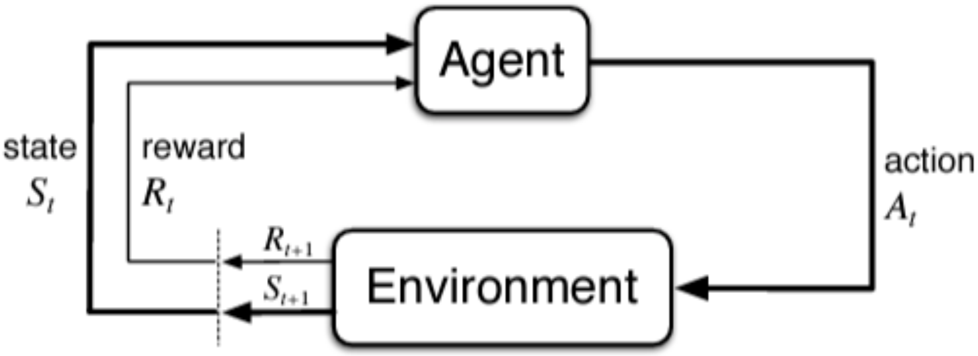


Figure 1.1: The agent-environment interaction in an MDP framework. At each time step t, the agent interacts with its environment by taking an action A_t based on the current policy it follows. After transitioning from state S_t to state S_{t+1}, the agent obtains feedback from the environment with (1) a reward signal R_{t+1}, if reward there is. It will be used to update the value of action A_t. And (2) sensory stimuli on the state S_{t+1} it is now in. Reproduced from Sutton et al. (1998) [5].

updated considering newest experience but also previous value estimates. More formally, the value associated to each state $V(s_t)$ is defined as the discounted sum of future rewards, and can be written recursively using Bellman's equation [19]:

$$\begin{aligned} V(s_t) &= \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r(\tau) \\ &= r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) + ... \\ &= r(t) + \gamma V(s_{t+1}) \end{aligned} \tag{1.2}$$

with s_t the state at time t, $r(t)$ the reward at time t and γ the discount factor. During learning, both side of equation 1.2 are estimate values of s_t. Due to the extra reward $r(t)$, the right-hand side is a more accurate estimate. The discrepancy between left-hand and right-hand side values or, more generally, between predictions at consecutive time points, is defined as the TD error δ [21]:

$$\delta(t) = r(t) + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t) \tag{1.3}$$

Finally the agent can update predicted value for state s_t, using TD error according to the following update rule:

$$\hat{V}(s_{t+1}) = \hat{V}(s_t) + \alpha\delta(t) \tag{1.4}$$

with α the learning rate ($0 < \alpha < 1$).

1.2 The RPE hypothesis

The *RPE hypothesis of DA neuron activity* proposes that the mesencephalic DA system distributes a signal that represents TD error [22].

The Rescorla-Wagner (R-W) model for classical Pavlovian conditioning - i.e. the association of a conditioned stimulus (CS), or cue, to a unconditioned stimulus (US), or reward - first proposed a link between learning and the changes in associative strength between the CS and US [24]. It hypothesizes that those changes are directly proportional to the discrepancy between the predicted and received US level, such that $\Delta\hat{V} = \alpha(V - \hat{V})$, with $\hat{V}$ and V, respectively the estimated and actual US, and α the animal's learning rate ($0 \leq \alpha \leq 1$). Experimental results collected by Schultz and colleagues show that the R-W model can predict accurately a smaller error signal at time of a predicted reward (Fig. 1.2.a, Top and Middle) [2]. However, it lacks a timing mechanism within the trial. Therefore, it cannot capture the cue response or the time-locked negative signal at the expected reward delivery time on reward omission (Fig. 1.2.a, Bottom). It was proposed that DA neurons convey TD error signals instead [2].

Indeed, a striking correspondence between error signals encoded by the DA neurons firing rate in the midbrain and the ones predicted by TD learning was found [25][23][26]. The neural basis of this TD error signaling was then further elucidated and confirmed through experiments on the firing activity of DA neurons in the ventral tegmental area (VTA) and medial substantia nigra pars compacta (SNc) in rodents [27][28][29][30][31] and humans, looking at the blood oxygen level-dependent (BOLD) response [32]. RPE signal encoding through phasic DA release in the nucleus accumbens was also assessed [33][34].

The original model (Schultz et al., 1997) used state features that constitute a complete serial compound (CSC) [2]. A stimulus is represented as a sequence of sub-states, tracking elapsed time relative to observable stimuli (Fig. 1.3, Left). The CSC representation considers that an animal in its environment experiences a sequence of states through time. Each state s_i is represented with a feature $x_i(t)$, containing description of the environment such as the cue onset or reward time. The value function estimate is then modeled as a linear combination of these features:

$$\hat{V}(t) = \sum_i w_i x_i(t) = w(t)^T x(t) \tag{1.5}$$

with $w(t)$, the vector of associative strengths w_i associated to each state s_i at time t. We can now express the weights update replacing value estimate with its linear combination from 1.2:

$$w(t+1) = w(t) + \alpha\delta(t)x(t) \tag{1.6}$$

with the TD error δ defined in Equation 1.4.

However, a CSC TD model cannot provide a complete account of timing. It responds equally to

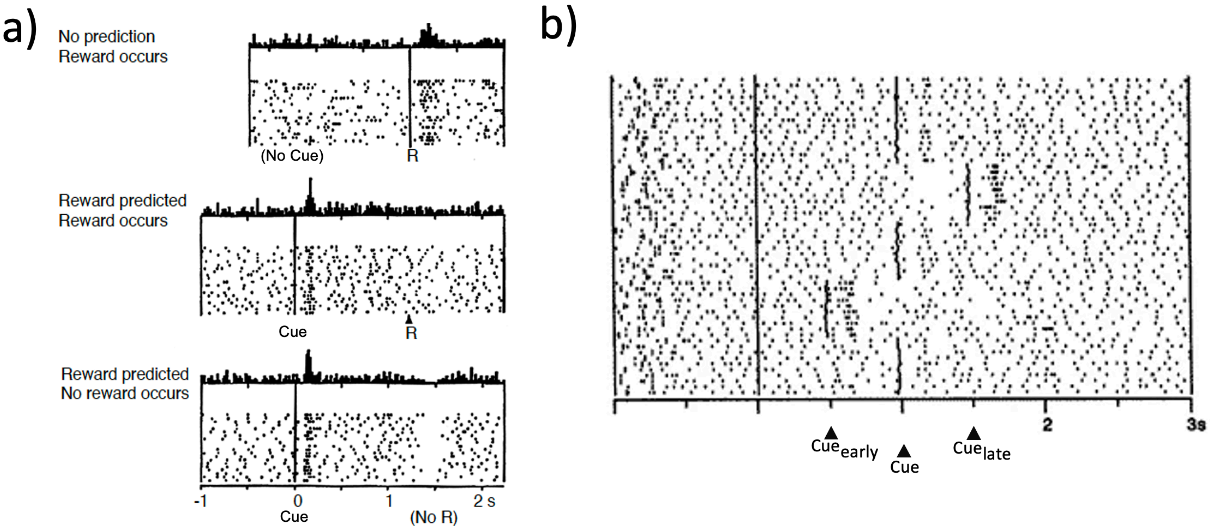


Figure 1.2: The reward prediction error (RPE) hypothesis. a. Firing pattern of a midbrain dopaminergic neuron during a classical Pavlovian conditioning task. **Top:** Before learning, no conditioned stimulus (CS) and absence of prediction. DA neurons fire for a positive error in the prediction of reward (unexpected reward). **Middle:** After learning, the CS (a cue) predicts an unconditioned stimulus (US) (a reward). DA neurons are expecting a reward, and the reward occurs. No error in the prediction of reward. **Bottom:** After learning, the cue predicts a reward. DA neurons are expecting a reward, but the reward is not delivered. The depression in the activity of the DA neurons at the time of the usual reward indicates an internal time representation of the reward. Original sequences of trials is plotted from top to bottom for the 3 peristimulus time histogram (PSTH). Adapted from Schultz et al. (1997) [2] **b.** Effect of reward timing during familiar trials. Following a cue - here being a correct response to an image selection task - the reward delay is either 0.5s (early), 1.0s (expected) or 1.5s (delayed). DA neurons activity is depressed when reward is delayed while it increases at the new delayed reward timing and at the earlier reward timing. Adapted from Hollerman & Schultz (1998) [23].

each sub-state marking the post-cue time steps, irrespective of the delay time from cue to reward. The microstimuli TD model is an alternative state features representation that swaps the CSC representation for Gaussian distributions, whose widths increase over elapsed time (Fig. 1.3, Right) [12]. Hence, the variance of timing estimation increasing linearly with elapsed time [35]. This model is consequently a temporally-diffused version of the discrete time markers in the CSC. It makes it more flexible to variable post-stimulus intervals and able to explain common observation in DA neurons recordings that were inconsistent with the CSC TD model [13].

The MDP framework and TD learning model as defined above assume a *stimulus-bound* state representation in which all relevant information is available to the agent. This is rarely the case in realistic natural scenarios. In 1998, Hollerman and Schultz trained monkeys on a cue-reward pairing with constant delay time. They actually observed that, when reward was given earlier than predicted, DA neurons activity response would be larger while there was no omission response at the time of the usual reward [23] (Fig. 1.2.b.). This does not fit with either of the CSC or microstimuli TD learning models. Consequently, a different framework in which the TD model features themselves represent the inferred state of the environment was proposed [8].

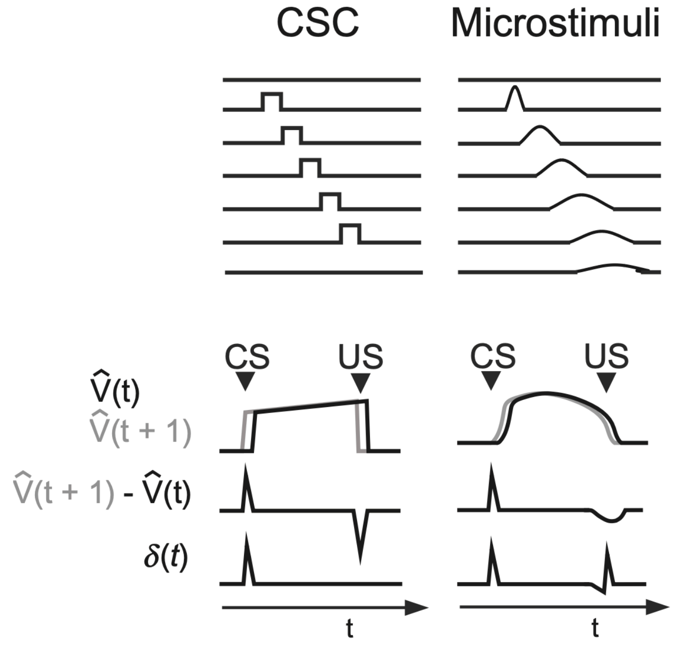


Figure 1.3: The CSC versus microstimuli state features representations. Left: The CSC features representation divides the post-stimulus time steps into a sequence of equally dispersed sub-states. **Rigth:** The microstimuli features representation is a temporally smeared version of the CSC, swapping the discrete time markers to Gaussian distributions, whose widths increase over elapsed time. For both representations, after the stimulus, only one of the sub-states becomes active at a time and the value function estimate is a linear combination of the stimulus features of each sub-state. Reproduced from Starkweather and Uchida (2020) [36].

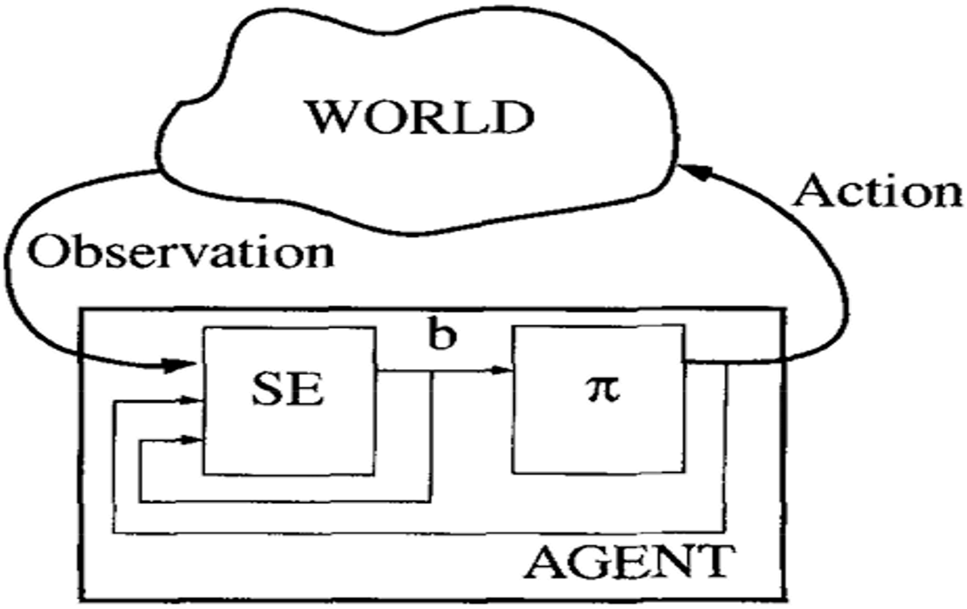


Figure 1.4: The agent-environment interaction in a POMDP framework. The agent can be decomposed into two parts. A state estimator (SE) is in charge of updating the belief state based on the last action, current observation and previous belief state. A policy π is, as in the MDP, responsible for generating action, but as a function of the belief state rather than states. Reproduced from Kaelbling et al. (1998) [6].

1.3 The partially-observable Markov decision process (POMDP) and belief state

When an agent has an incomplete knowledge of the world, it cannot learn the optimal policy. Instead, it must estimate the current *hidden state* by combining current noisy observations with predictions from previous state estimates and actions, using an internal model [6]. Consequently, we can differentiate two decision-making strategies. The *stimulus-bound* decision-making strategy, described above, is deterministic. It relies only on current perceptual information and observable states to compute value. State representation constitutes either of CSC or microtimuli. Alternatively, the *inference-based* decision-making strategy represents the states by inferring the posterior probability distribution over possible states, forming the belief state [8]. Value is no longer proportional to the weights associated to each state, such as in Equation 1.2. Instead, it is equal to the weight associated to each state scaled by the probability that the agent believes it is in that state, over all the states:

$$\hat{V}(t) = \sum_i w_i b_i(t) = w(t)^T b(t) \tag{1.7}$$

with $b(t)$, the belief state (i.e. the probability to be in each state) at time t replacing $x(t)$, the feature vector describing the environment from Equation 1.5. The other elements are similar to those in Equation 1.5, used to compute value estimates from a CSC features representation. The critical difference between the two strategies is that in the stimulus-bound representations, only one sub-state x_i becomes active (non-null) at a given time in $x(t)$. In the inference-based model, the b_i constituting $b(t)$ form a distribution over all states. Thus, they can all be non-null at a given time step, depending on how certain the agent is at that time step.

The partially-observable Markov decision process (POMDP) framework (Fig. 1.4) formally defines the elements needed to the inference-based strategy [6]. It can be described as a tuple (S, A, T, R, Ω, O) where:

- S, A, T and R describe an MDP framework, as in Section 1.1;
- Ω is a finite *set of observations* that the agent can experience from its environment;
- O is the *observation function* so that $O_a(s, s')$ is the probability of making observation o after transitioning from s to s', taking action a .

Building on that POMDP framework, several models for neural implementation, using TD error for learning were theoretically proposed [9][8]. They present the same core elements:

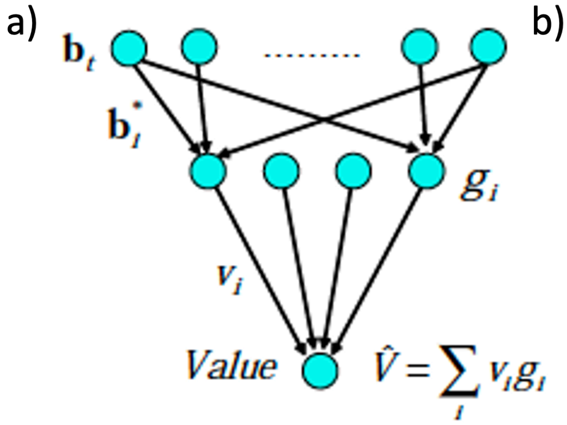
- A policy π, as a function of belief state, that depends on the task and define the rules for action selection.
- A state estimation system, inferring the world's state using noisy observations and previous state. It relies on the - critical - notion of belief state. Considering discrete states, the belief state is modelled as a vector b_t, whose size is the number of states. Its i^{th} component is the posterior probability of state i: $b_t(i) = P(s_t = i|o_t, a_{t-1}, o_{t-1}, ..., a_0, o_0)$. It is updated recursively using Bayes' rule, converting prior beliefs into posterior beliefs:

$$\begin{aligned} b_t(i) &\propto P(o_t|s_t = i, a_{t-1}, o_{t-1}, ..., a_0, o_0)P(s_t = i|a_{t-1}, o_{t-1}, ..., a_0, o_0) \\ &\propto P(o_t|s_t = i)\sum_j P(s_t = i|s_{t-1} = j, a_{t-1})P(s_{t-1} = j|o_{t-1}, a_{t-2}, ..., a_0, o_0) \\ &\propto P(o_t|s_t = i)\sum_j T(j, a_{t-1}, i)b_{t-1}(j) \end{aligned} \tag{1.8}$$

  with T, the state-transition function. The Markov property was assumed, so that $P(o_t|a_{t-1}, o_{t-1}, ..., a_0, o_0)$ is constant and $P(o_t|s_t = i, a_{t-1}, o_{t-1}, ..., a_0, o_0)$ is proportional to $P(o_t|s_t = i)$. Various models for neural implementation of Bayesian inference attempt to explain how the belief state could be computed through the neural circuitry [37]. They rely on a probabilistic graphical model framework [38] and a message-passing algorithm for inference, such as belief propagation [39].

- A value prediction system, that uses TD error signals to map estimated state representations to expected rewards. For a fixed policy π, the value estimate is defined as the expected sum of rewards, starting from the current belief state. It can be written recursively from Bellman's equation [19], and similarly to equation 1.2:

$$\begin{aligned} V^\pi(b) &= E_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|b_t = b] \\ &= E_\pi[r_{t+1} + \gamma V^\pi(b_{t+1})|b_t = b] \end{aligned} \tag{1.9}$$



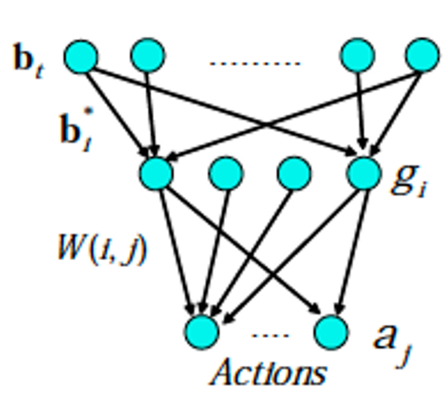


Figure 1.5: Neural implementation of inference-based value estimate and action selection models. a. The value is estimated through a three-layers network (top to bottom). The input layer receives the belief state b_t. The hidden layer represents a set of Gaussian basis function so that each neuron i of the layer is activated in proportion to how close the current input belief state is to its preferred belief point b_i^*. The firing rate of the i^{th} hidden neuron is such that $g_i(b_t) = e^{-\|b_t - b_i^*\|^2/\sigma^2}$. Making the parallel with computational theories, one can see the preferred belief points as synaptic weights from input to hidden layer, using an exponential activation function. Finally, the output of the network is the estimate of the value, given by $\hat{V}(b_t) = \sum_i v_i g_i(b_t)$, with v_i the synaptic weights from the hidden layer to the single output neuron. **b.** The input layer and synaptic weights b_i^* of the action selection network are the same as in the value estimation network (top to bottom). The output layer represents the set of K possible actions, one of which is being selected in a probabilistic manner at each time step so that $P(a_j|b_t) = exp^{\sum_i g_i(b_t) w(i,j)/\lambda}/Z$, with $W(i,j)$ the synaptic weight from hidden neuron i to output neuron j and λ the coefficient qualifying the exploitation-exploration trade-off - 0 rendering the policy fully-exploitative and larger values allowing more action exploration. Reproduced from Rao (2010) [9].

which suggests an online learning rule for value estimation by minimizing the error function:

$$e = (V^{\pi}(b_t) - (r_{t+1} + \gamma V_{\pi}(b_{t+1})))^2 \tag{1.10}$$

with e being the squared TD error presented in equation 1.4 [21]. Rao (2010) proposed a value estimator neural network (Fig. 1.5.a.). The synaptic weights b_i^* and v_i in this network are learned through a TD learning model based on belief state [9].

- An action selection system, that uses the trained value estimator outputs to learn an action selection policy. Rao developed an action selection neural network using the previously defined value predictor neural network (Fig. 1.4.b.).

Following those theoretical proposals, the Uchida/Gershman labs developed two experimental paradigms in mice to study how and, if so, where belief states are provided as inputs to the DA system. Babayan and colleagues (2018) investigated how hidden state inference affects DA neurons activity responses in the VTA and behavior across trials using the *reward inference* paradigm. Mice are trained on two states distinguished by their reward magnitude: *small* and *large* rewards. Then, new intermediate rewards are introduced, with the goal of inducing ambiguity onto which state they are indicative of [15]. Results were in concordance with predictions from a TD learning model using belief state. The response of DA neurons was found to be a non-monotonic function of reward size. When a small intermediate reward, intuitively assigned to the *small reward* state and perceived as *better than expected*, was delivered, DA neurons activity increased. Large intermediate rewards, that should be assigned to the *large reward* state and hence perceived as *worse than expected*, provoked a decrease in the DA neurons activity.

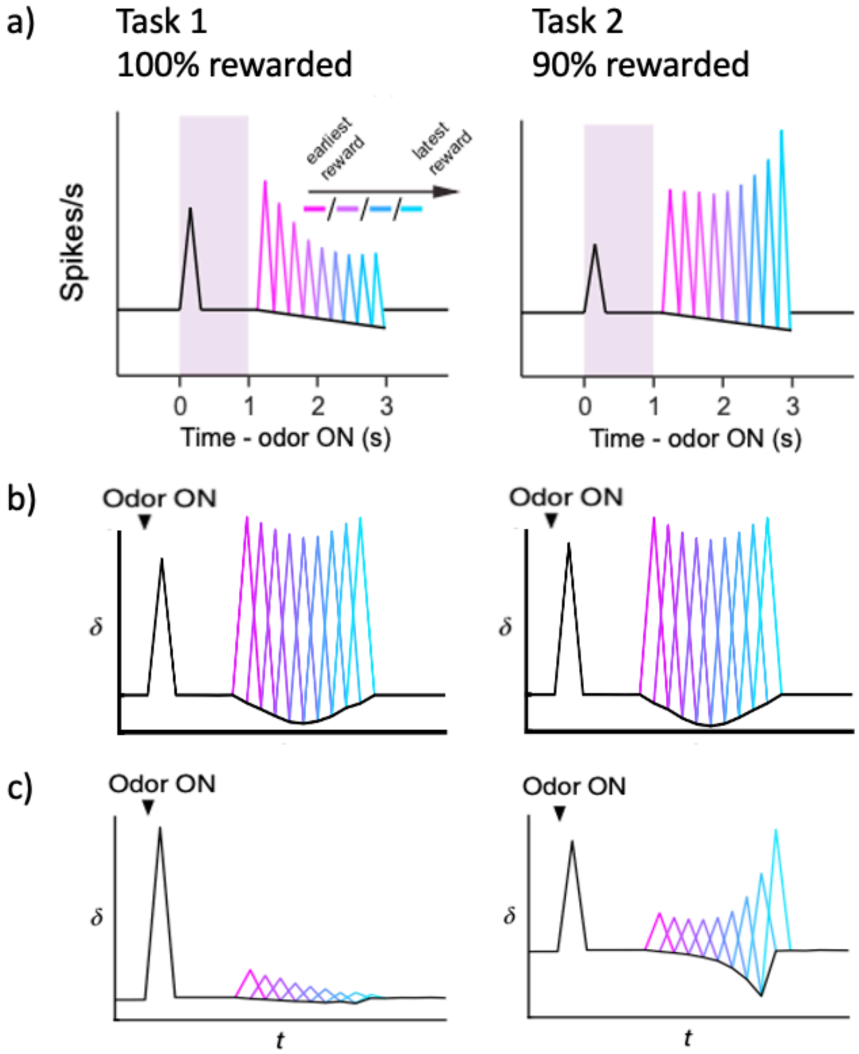


Figure 1.6: Neural activity and RPEs patterns predicted by a TD model with CSC and belief state representations on the variable reward delay paradigm a. Average non-normalized peristimulus time histogram (PSTH) for 30 DA neurons recorded during trials for odor A (randomized Gaussian distributed reward) in task 1 (left) and task 2 (right). Shaded rectangle indicates 'odor ON'. RPEs when reward delay varies decrease as a function of time in task 1, whereas they increase in task 2. **b.** Predicted RPEs pattern for TD model with CSC state representation resemble the flipped reward time probability distribution (Gaussian distribution) for both tasks. Plots are averaged from ten simulations of 5,000 trials each. **c.** Predicted RPEs pattern for TD model with belief state representation captures the opposing post-reward firing rate patterns and negative temporal modulation of pre-reward firing rate observable in empirical data for both tasks. Plots are averaged from one simulation of 3,000 trials for task 1 and 2,700 trials for task 2. Adapted from Starkweather et al. (2017) [14].

The second paradigm, the *variable reward delay task* (Starkweather et al., 2017, 2018 [14][16]), was designed to investigate the dynamics of belief state within a trial. It consists of two Pavlovian conditioning tasks that differed only on whether or not reward was always delivered following a given stimulus. Reward was always presented in task 1 while there was a 10% probability of reward omission in task 2 (Fig. 2.1 for the detailed paradigm). We will further refer to the interval between the odor onset and the reward onset as inter-stimulus interval (ISI) and the interval between the reward onset of a trial and the odor onset of the next trial as inter-trial interval (ITI). Moreover, early ITI corresponds to the ITI period following reward, while late ITI, precedes the odor onset. DA neurons activity responses recorded in the VTA decreased as a function of the ISI in task 1 while they increased in task 2 (Fig. 1.6.a.). TD model using a CSC state representation could not explain the empirical data (predicted patterns in Fig. 1.6.b. differ from empirical data in Fig. 1.6.a.). Alternatively,

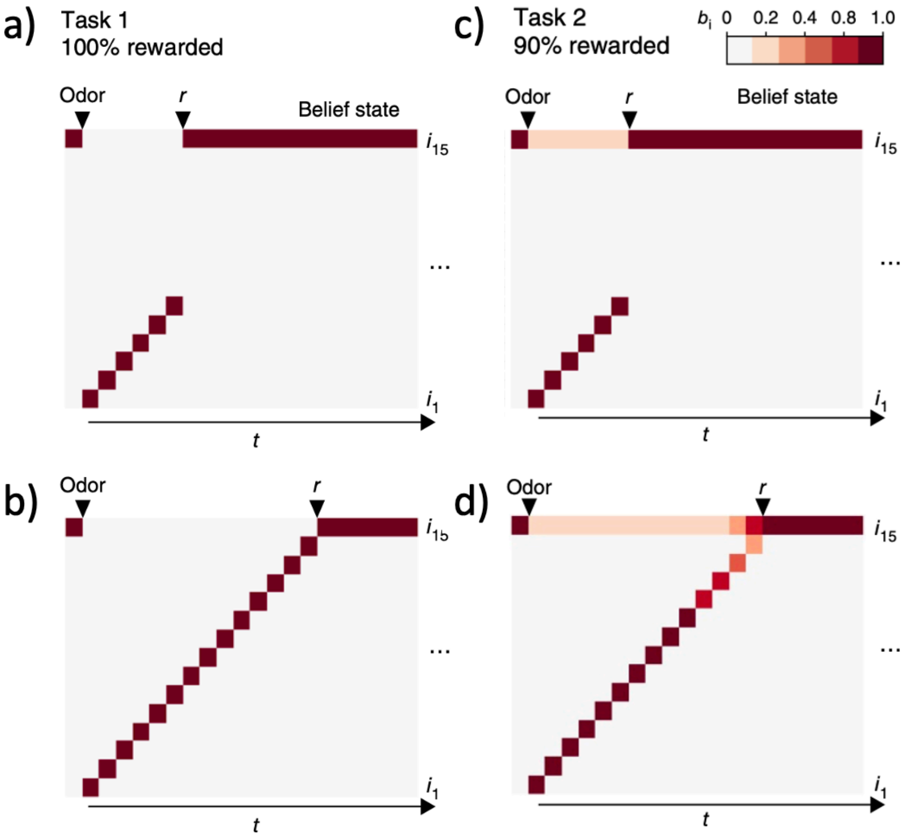


Figure 1.7: Representation of the belief state in the variable reward delay paradigm. (a,b) Belief state from one representative simulation for task 1 (n = 3,000 simulated rewarded trials), for earliest **a.** and latest **b.** rewards in task 1. As time elapses following odor onset in task 1, the belief state proceeds through ISI sub-states i_1–i_{14} by sequentially assigning a 100% probability to each sub-state. **(c,d)** Belief state from one representative simulation for task 2 (n = 2,700 simulated rewarded trials). As time elapses following odor onset in task 2, the belief state comprises a probability distribution that gradually decreases for ISI sub-states i_1–i_{14} while gradually increasing for the ITI sub-state i_{15}. Adapted from Starkweather et al. 2017 [14].

a TD model relying on belief state representations, tracking posterior distribution over temporal sub-states, analogously to the sub-states of the CSC model (Fig. 1.7) was designed. Its predictions captured the RPEs modulations from the experimental data well (Fig. 1.6c.).

Taken together, these studies, as well as previous and more recent empirical studies, present complementary evidence that, when a task includes uncertainty, the brain computes RPEs over the belief state, assuming a limited set of discrete states [40][10][41]. Arising questions are then *where* and *how* the brain is estimating value based on beliefs. Here we will consider two possible hypotheses as to *how* beliefs are encoded. First, the brain could estimate value using a two-stage model. The brain estimates beliefs using observations, and then learns value using those beliefs, either in the same or in different regions [37], similarly to the POMDP framework presented by Kaelbling and colleagues (1998) (Fig. 1.4; [6]). Alternatively, the brain could also estimate value directly from the observations. By only tracking the subspace of beliefs that are relevant to value prediction, it could use a compressed hidden state that is sufficient for estimating the value.

1.4 Belief state, mPFC and OFC

Regarding *where* the belief state might be computed, literature mostly agrees on the prefrontal cortex. Several studies targeted the mPFC and the OFC [16][42].

A follow-up study on the variable reward delay task used reversible chemogenetic inhibition. κ-opioid receptors (KOR)-DREADD (KORD) in the mPFC were inhibited using a salvinorin B (SalB) injection [43]. Dopaminergic signaling in task 2 but not in task 1 was affected [16]. The corresponding RPEs patterns resemble signaling observed in the neural activity in task 1 and the flipped reward distribution pattern predicted by the stimulus-bound state representation model (Fig. 1.6.b.). mPFC inactivation froze inferred probability distribution over states that should have evolved over time. This suggests that the mPFC is required for hidden state inference. However, the ISI and ITI were still discriminated through observable cues such as reward. A basic representation of the observable state space remains intact, hinting that it is computed in a different region.

Another line of work suggests that activity in the OFC is at the origin of such a state space representation [42]. If this is the case, OFC would learn to represent the hidden state space over which the mPFC would compute a probability distribution, i.e. the belief state [16]. mPFC could also encode state representation but at a larger scale, for macro-states, i.e. the probability to be in the ITI or ISI. The OFC would then encode the detailed beliefs of the micro-states in the ISI, similarly to the sub-states in the CSC features representation. A brain region should satisfy two conditions to encode state representation: (1) the representation condition, by representing sufficient information of the world, meaning all variables relevant to a given task, and (2) the specificity condition, by representing relevant information of the world, meaning variables irrelevant to the task are not encoded. It also must be able to deal with partial observability, by combining necessary unobservables, such as past experiences. Consequently, such a brain region would need to have access to both sensory cortices and brain areas relevant to memory [42].

The OFC is a large cortical area located at the most ventral surface of the pre-frontal cortex (PFC), above the orbit of the eyes [18]. Interestingly, it is closely connected to all five sensory areas as well as to other parts of the frontal cortex and learning and memory structures such as striatum amygdala and hippocampus [44][45]. Hence, on solely anatomical grounds, the OFC is a candidate of choice for state representation. In addition, it has been found to be implicated in a wide range of tasks requiring decision-making and involving hidden states. The OFC encodes variables such as uncertainty [46], spatial location [47], taste [48], odor [49], cue-outcome associations [50][51][52] and history of previous outcomes and decisions [53]. Consequently, it was proposed that the OFC represents partially-observable states of the world. It could encode the current state in a cognitive map of task space by providing an abstraction of the currently available information, necessary for the task [42].

Coming back to the necessary conditions to encode state representation, the *shift-stay* paradigm asserts the representation condition. In that paradigm, monkeys choose between two options. A cue instructs them as to whether the rewarded response is to *stay* with their last choice or to *switch* to the other option. The two states are consequently a combination of the last choice and the cue. Neural correlates of both these variables were actually found in the OFC [54]. Similarly other studies showed results consistent with these observations [55][56]. For the specificity condition, changes in the relevance of task variables should lead to changes in the OFC neurons firing rate. To assess that, two tasks with similar stimuli - eight odors - but different underlying states were compared. In the first one, four of the odors predicted a reward, while the four others did not [49]. In the second, reward occurrence in the current trial depends on whether the odor is similar to the one from the previous trial [57]. For both tasks, odor is relevant to performance. However, identity of the odor is critical only in the first one, while the second relies on the matching of consecutive odors. The results showed that around 77% of the OFC neurons were odor selective when odor identity was relevant, in the first task. On the second task, only 15% of them were odor selective, while 63% encoded whether the odor was a match or no [42]. More recent paradigms have found OFC implicated in state-inference processes [41]. OFC inactivation also reverted animals from an inference-based to a stimulus-bound decision-strategy but did not impact performances when variables were fully observable [58].

Hence, mPFC and OFC seem to be at the origin of belief state computation. In particular, OFC neurons encode necessary and sufficient features of the state representations, while being conveniently connected to areas potentially required for such a computation. However, the precise neural basis of how all the task features are combined to form belief states still needs to be investigated.

1.5 Belief state, population analysis and attractor-like dynamics

Individual neurons in the OFC can encode task variables, and more precisely environmental stimuli required for belief computation. Thus, belief state could be encoded at the population level, by combining individual neuron activity, encoding the features of such states.

Single-neuron electrophysiology has provided a revolutionary framework to some of the biggest neurosciences discoveries of the last century, some even leading to Nobel prizes [59]. A common approach in single-neuron analysis is to find a mapping between activity of individual neurons and stimuli from the external world, behaviors or internal functions by considering the mean neuronal response as a function of those variables. Nowadays, with the advent of experimental tools to record hundreds of neurons at the same time in behaving animals (e.g. Neuropixels, high-density neuronal recording probes [60]) and powerful machine learning tools to analyze, reduce and interpret high-dimensional data, neural population level analysis is pervading systems neuroscience [61].

Analytic tools can be regrouped into *encoder* and *decoder* models [62]. Encoders use stimulus features to model brain activity. They describe how dynamic stimulus features are encoded into patterns of neural activity so that $activity = Sw + \epsilon$, where S is the stimulus matrix with each row corresponding to a time point response and each column corresponding to the feature values at that time point, w is a vector of model weights and ϵ is a vector of random noise at each time point.

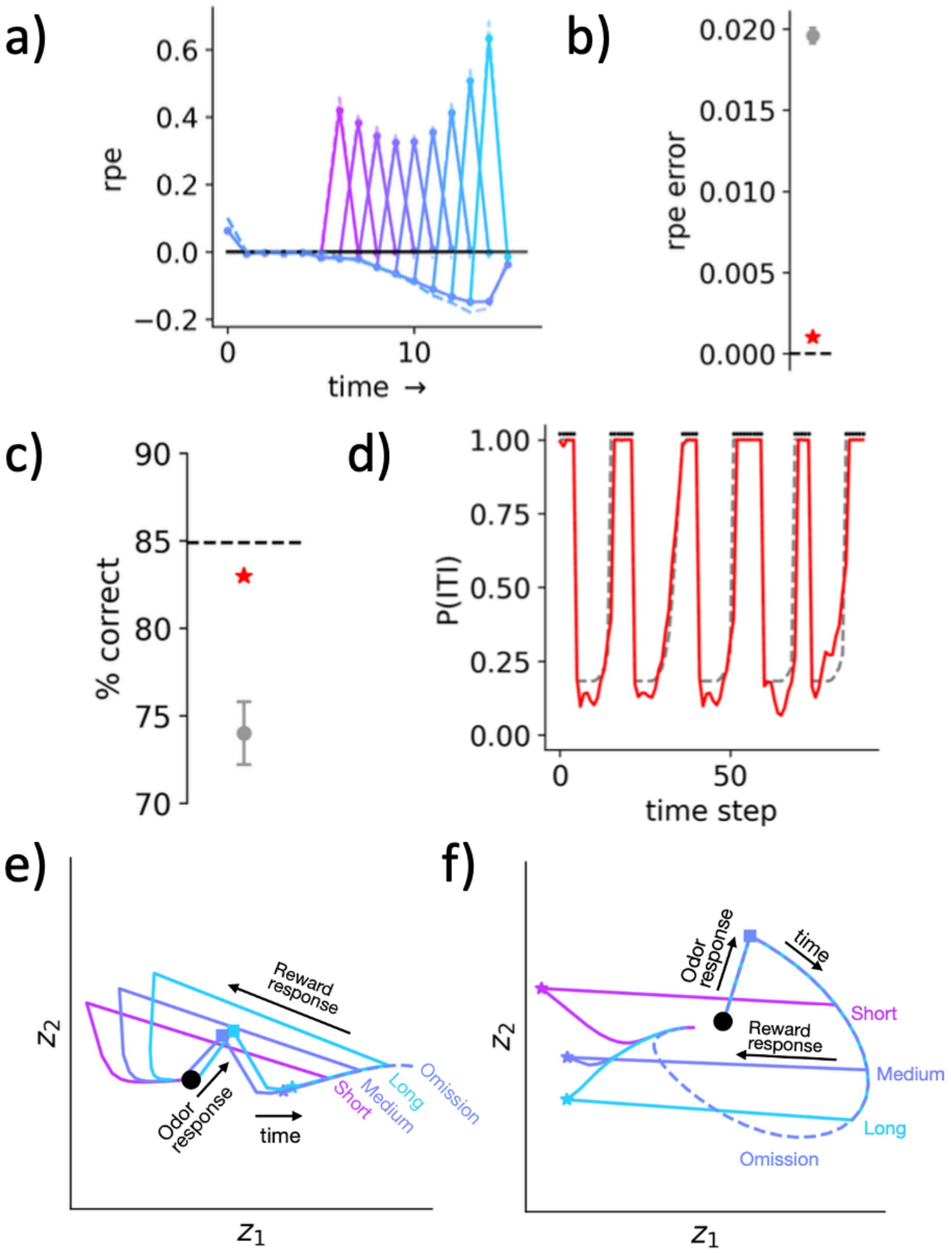


Figure 1.8: Performances comparison between an MDP and POMDP model using TD learning and GRU network trained on the task All frameworks were evaluated for odor A trials only, on task 2. **(a,b) recurrent neural network (RNN) trained on observation predicts the same RPE patterns as the POMDP estimations from [14]; a.** Estimates of the RPEs for the RNN (solid lines) and POMDP (dashed lines) frameworks, for odor A trials, with different reward timings, in task 2 (90% rewarded); and **b.** Averaged mean squared error (MSE) between estimates of RPEs of POMDP and RNN (red star) and of POMDP and untrained RNNs (grey dot and whiskers; mean ± standard error, 100 random RNNs). **c,d. RNN activity resembles beliefs; c.** Linear decoders trained on RNN activity (red star) or beliefs (black dashed line) to estimate in which of 15 different states the experiment was at each time step in task 2 (90% rewarded). Decoder performances are quantified as percent correct. Grey dot and whiskers indicate the decoder performances when untrained (mean ± standard error, 10 random RNNs); **d.** $P(ITI)$ estimate over 5 trials, using the decoder trained on RNN activity (red) and beliefs (grey dashed) compared to true state (black dots indicates when true state is ITI). **e,f.** Hidden activations activity of the RNN for tasks 1 and 2 on odor A encode two fixed points corresponding to the macro-states, namely ISI andITI. The dynamic is different depending on the reward timing; and **e.** in task 1, the activity trajectory cannot return to ITI (left of the odor response) without getting a reward; while **f.** in task 2, activity returns to the ITI through a circular trajectory if reward is omitted while reward induces a rapid return to the ITI region of the hidden activation space. Reproduced from Hennig, unpublished work.

On the other hand, decoders use neural features to generate a stimulus output. They allow to infer the stimulus or experimental properties that were most likely present at each time point so that $s = Xw + \epsilon$, where s is the vector of stimulus feature values over time, X is the activity matrix with each row a time point and each column a neural feature, w is a vector of model weights and ϵ a vector of random noise at each time point.

Tools to reduce and interpret neural population data have already proven their utility. Successful studies on decision-making unraveled the neural basis of changes of mind [63], subjective decision [64] or anticipatory choices [65] using those techniques. Population level analysis has the advantage over single-neuron approaches to allow the investigation of variables encoded at the population level rather than single-neuron level, for neuronal processing requiring neurons coordination [61]. For instance, combining both levels of investigation, it was found that a region could show stability at the population level for a variable despite unstable representation of the same variable at the single-neuron level [66][67].

Theoretically, RNN are considered to approximate dynamical systems [68]. Consequently, to investigate neural activity at the population level, it is common to compare the dynamics of an RNN to those of a biological neural network (i.e. the brain region of interest), trained on the same task. By comparing the dynamics of their neural activity, one can infer the mechanisms and representations at play in the biological circuit and elucidate its computational function. However, this requires an understanding of which elements are comparable and which are variable depending on the architecture of the network. It was found that the geometry of the neural representations in an RNN can be highly sensitive to the network's specific architecture and activation function. However, the topological structure of fixed-points and transitions between them are shown to be sensitive to the task but mostly universal across architectures [69]. Consequently, at the population level, one needs to carefully consider measures performed on representational geometry such as principal component analysis (PCA) or support vector machines (SVM). One should rather prefer analyses on the dynamics of the activity, such as fixed-points and their transitions.

Consequently, preliminary works at the Uchida/Gershmann labs (Yamaguchi et al., Hennig et al.; unpublished works) aimed at investigating belief computation at the population level using artificial neural network (ANN) modeling. A gated recurrent unit (GRU) network [70][71] consisting of three hidden units was trained and analyzed by Dr. Jay Hennig (Fig. 1.8), while Takahiro Yamaguch worked with a uniform 4-layers RNN using standard optimization in time such as feedback alignment [72] or nonlinear function approximation [73]. Both were trained on the variable reward delay task [14] to directly learn value by *only* minimizing TD error from observations, without the explicit objective of learning beliefs. Several observations were made:

- Outputs of the RNNs trained on task 1 and task 2 can recapitulate the RPEs observed in the empirical DA neuron firing and the TD model operating with belief state presented in [14] (Fig. 1.8.a. for the GRU network).

- Predictions of the macro-state (ISI or ITI), using the hidden activations of the GRU network, were as good as using the true beliefs (Fig. 1.8.(c-d)). This suggests that the hidden activity of the network naturally resembles true beliefs. Hence, while optimizing value, the model discovered an implicit belief-like representation of the task.

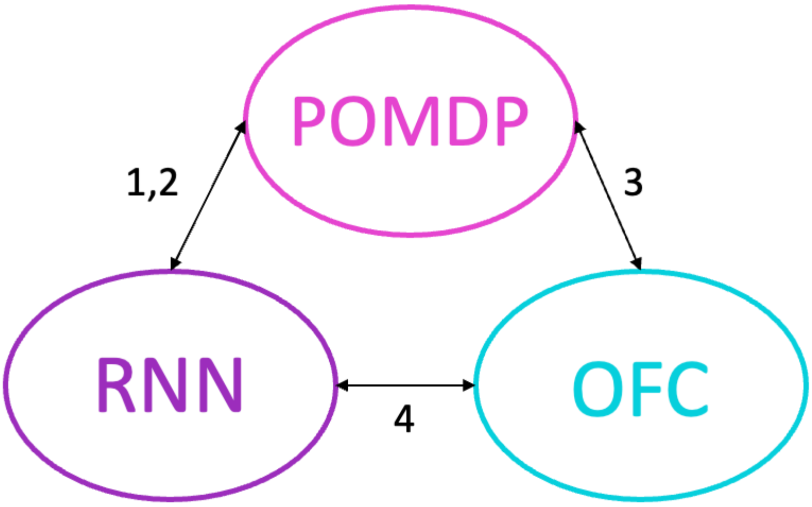


Figure 1.9: Connections between the three components estimating value on the task. Those connections need to be assessed to elucidate the neural basis of belief state computation in the brain. The components consist in the POMDP framework using belief-state TD learning, the GRU network estimating value and the brain (OFC). Connections are **(1)** The RNN model can recapitulate the estimated values and RPEs of the POMDP model; **(2)** To estimate value from observations, the RNN model learns a representation that resembles beliefs; **(3)** Neural activity in the OFC/mPFC resembles beliefs. **(4)** Neural activity resembles the RNN activity.

- Looking at the hidden activity of the GRU network, i.e. the two activations layers learned at each time point, the network developed attractor-like fixed-points, corresponding to respectively the ISI and ITI. In task 1 (100% reward; Fig. 1.8.e.), activity goes quickly from stable ITI to stable ISI state at cue onset and from stable ISI to stable ITI state at reward. In task 2 (90% reward; Fig. 1.8.f.), the ISI state is *leaky*: the activity gradually moves away from the ISI state, towards the ITI state as time elapsed. When reward is delivered, the activity jumps back to the ITI state.

While artificial, those results provide great insights into the dynamics of the activity that may underlie the representation of belief states in neural circuits. Figure 1.9 presents the connections between the components we will investigate and that are capable of estimating value on the task: the POMDP framework (belief-state TD model), the GRU network and the brain itself. Qualifying the nature of those connections better will help understand how beliefs are computed. We already have seen that the RNN can recapitulate the estimated values and RPEs of the POMDP model (Fig. 1.8.(a-b)). Moreover, the RNN model seems to internally learn a representation that resembles belief in order to estimate value from observations (Fig. 1.8.(c-d)). Consequently, remaining investigations concern linking neural activity to beliefs and RNN activity.

To address these, we seek to analyze neural activity both at the single-neuron and population level. We will train a linear classifier to recognize states based on the neural activity, similarly to what was done with the hidden activations of the GRU network and beliefs of the POMDP. We will also investigate dynamical trajectories of the neural activity and link it to the fixed-points observed in the dynamics of the hidden activations of the GRU network.

Chapter 2

Methods

Data collection and preprocessing was performed by student Sandra Romero Pinto (Uchida lab) and the GRU network design and analysis were completed by Dr. Jay Hennig (Gershman lab).

2.1 Behavioral paradigm and data collection

Data collected for this project consist of neuronal activity in different regions of the brain, recorded using high-density neuronal recording probes (Neuropixels, [60]), while mice were performing the variable reward delay task [14].

2.1.1 Mice

We used adult male and female mice (wild-type C57/BL6J). The mice were housed on a 12-h dark:12-h light cycle (dark from 7a.m. to 7p.m.). We trained mice on the behavioral task at approximately the same time each day. All experiments were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals [74] and were approved by the Harvard Institutional Animal Care and Use Committee.

2.1.2 Surgery

We performed all surgeries under aseptic conditions with mice that were under isoflurane (1–2% at 0.5–1.0 liter/min) anesthesia. Analgesic (buprenorphine, 0.1 mg per kg body weight; by intraperitoneal injection) was administered pre-operatively and at 12-h checkpoints post-operatively. We performed two surgical procedures. In the first surgery, we implanted a steel head plate. Prior to surgery, the head plate is pumiced to increase surface area and adhesion strength. It is fixed to the skull using adhesive luting dental cement (CB-Metabond) applied to all surfaces of the head plate-skull interface. In addition, a small craniotomy ground pin is inserted into the cerebellum through a small (0.15 mm) craniotomy. During this surgery, a portion of the skull is left exposed, and it is marked with a surgical pen on the stereotactic coordinates of future craniotomies, above the brain target region to record from. Nine markings were performed in each hemisphere at the level of: 0.6, 1.0, 1.6 mm lateral and 3.2, 2.6, 2mm ventral from bregma. The exposed portion of the skull is then covered with silicon sealant (KwikCast, World Precision Instruments). After the surgery, mice were allowed to recover for at least 5 days with free access to water before water restriction and behavioral training.

In the second surgery, a craniotomy is performed on the animal to create a small window (0.6 mm) in the skull to insert the Neuropixels probe in the brain. As with the head plate surgery, the craniotomy is carried out under aseptic conditions. Coordinates of the craniotomy are indicated by the previously made markings. This surgery is performed the night before each recording session. The exposed portion of the skull including the craniotomy, is then covered with silicon sealant (KwikCast, World Precision Instruments).

2.1.3 Behavioral training

After 1 week of post-surgical recovery, we water-restricted mice in their cages. Weight was maintained above 85% of the body weight prior to the water restriction. We habituated and briefly head-restrained mice for 2–3 days before training. Animals were then trained on the recording apparatus for at least two weeks or until demonstration of full task comprehension. The task structure was kept the same during training and recording sessions. odors were delivered to mice with a valve driver (Sanworks) using solenoid valves (Lee company). Each odor-emitting compound was dissolved in mineral oil at 1/10 dilution. 30μL of diluted odor-emitting compound was placed on a glass fiber filter paper. Filtered air was run through the filter paper to produce a total flow rate of 1 L/min. Odor-emitting compounds included isoamyl acetate, (+)-carvone, 1-hexanol, p-cymene, ethyl butyrate, 1-butanol, limonene, dimethoxybenzene, caproic acid, 4-heptanone and eugenol. The combination of these odors differed for different mice. Licks were automatically detected by measuring breaks in an infrared beam that was placed in front of the water spout. The paradigm used, the *variable reward delay task* paradigm, was developed by the Uchida/Gershman labs and consisting of two Pavlovian conditioning tasks (Fig. 2.1.(a-d)) [14][16]. In the task, rewarded odor A trials consisted of a 1-s odor presentation followed by a delay sampled from a Gaussian distribution defined over nine points (1.2s 1.4s, 1.6s, 1.8s, 2.0s, 2.2s, 2.4s, 2.6s and 2.8s; mean = 2s; s.d. = 0.5s) prior to reward delivery and defined as the ISI; rewarded odor B and odor C trials consisted of a 1-s odor presentation followed by either a 1.2-s or 2.8-s from odor onset prior to reward delivery, respectively; odor D trials were always unrewarded. In task 1, reward was delivered in 100% of trials (Fig. 2.1.a.), making the task state fully observable (Fig. 2.1.b.) and the belief state fixed at 100% of chance of being in the ISI after cue onset and until reward delivery (Fig. 2.1.c.). In task 2, reward was delivered in 90% of trials (Fig. 2.1.d.), making the task state hidden, with the probability of being in the ISI after cue onset progressively decreasing as time elapses and until reward delivery or expected reward time step passes. For all tasks, reward size was kept constant at 3μL. Trial type was drawn pseudo-randomly from a scrambled array of trial types, to keep the proportion of trial types constant between sessions. The ITI between trials was drawn from an exponential distribution (mean=12–14s, min=4s) to ensure a flat hazard function. Mice performed between 100 and 300 trials per recording session.

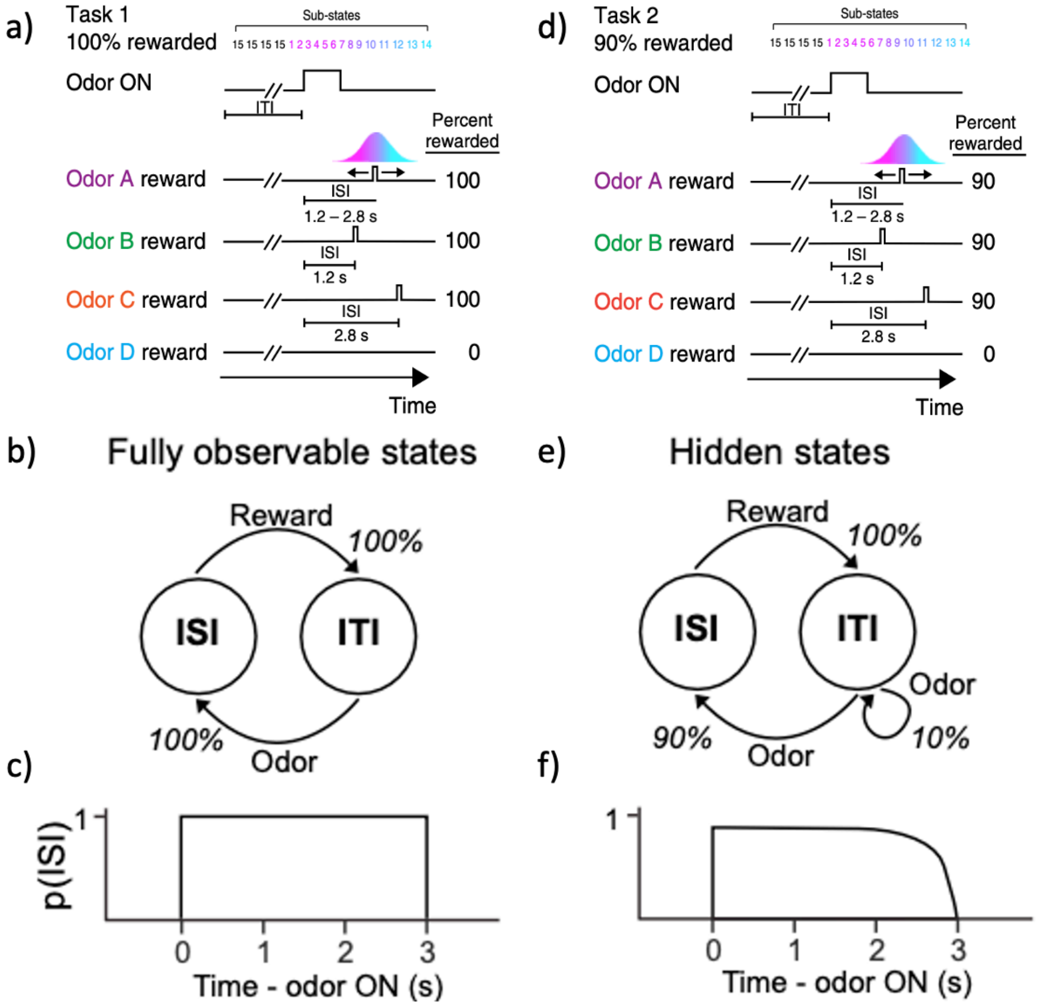


Figure 2.1: Variable reward delay paradigm For both task 1 and task 2, odor onset happens 2s after trial start. States consists of the ISI, the interval between the cue and the reward and ITI, the interval between the reward of a trial and the cue onset of the next one. Trials for odor A have a variable ISI, drawn from a discretized Gaussian distribution defined over nine time points between 3.2s and 4.8s after trial start. Trials for odors B and C have constant ISIs with reward timing at 3.2s and 4.8s respectively from trial start. Trials with odor D are non-rewarded. A session consists of 100 to 300 trials, randomly selected among the 4 odors. **a.** In task 1, rewarded odors forecasted a 100% chance of reward delivery, reward timing would vary; **b.** the task state (either ISI or ITI is fully observable because it is reliably signaled by sensory observations (cue and reward); **c.** belief state is fixed at 100% probability in the ISI state after cue onset. **d.** In task 2, rewarded odors forecasted a 90% chance of reward delivery; **e.** task state is initially fixed at a 90% probability in the ISI at cue onset with the probability progressively decreasing as time elapses and until either the reward happens or the expected reward timing is passed; **f.** belief state is initially fixed at 90% of probability and progressively decreases as time elapses to reach a probability that the trial will be unrewarded, allotting more probability to be in the ITI. Adapted from Starkweather et al. 2017, 2018 [14][16].

2.1.4 Electrophysiological recordings

The animal is placed in the recording rig and the artificial dura and silicone sealant are removed. Under a microscope, any emergent scar tissue is removed with sterilized precise forceps. The silicon probe (Neuropixels 1.0 phase 3A) is then dipped into one of the following dyes: 1,1′-Dioctadecyl-3,3,3′,3′-tetramethylindocarbocyanine perchlorate (DiI) or 3,3′-Dioctadecyloxacarbocyanine perchlorate (DiO), dissolved in ethanol. This was done to ensure recovery of the real coordinates at which the recordings were performed. The probe was then advanced into the brain at a controlled speed of

0.005 mm/sec using a motorized stage (Thorlabs) aligned to the dorso-ventral axis of the brain. The probe insertion was made up to a depth of 3.5 mm from the pia surface. The probe was allowed to settle 20 minutes before beginning the behavioral session and neural recordings. To prevent brain from drying, a sterile saline cortex buffer solution was applied on the craniotomy. The neural signals were recorded using an PXIe acquisition device (NI PXIe-1082 chassis, PXIe-8381 interface, National Instruments) and the SpikeGLX software package. Broadband signals were recorded continuously at 32 kHz. After recording, the probe was slowly retracted and the craniotomy resealed with sterile artificial dura and silicone sealant. Recordings were made one time per craniotomy, and lasted 10-20 days after the first craniotomy was performed or unit yield in superficial layers decays substantially due to degradation of tissue.

2.1.5 Neural signals processing

Raw data within the action potential band (soft high-pass filtered over 300 Hz) was denoised by common mode rejection (that is, subtracting the median across all channels), and spike-sorted using Kilosort2 [75]. This software extracts clusters of spike times, based on waveform features, spike activity across the recording sessions and spatial information of the probe (i.e., electrodes map). Curation of the clusters for further analysis was done using Phy [76]. This software was used to classify a cluster as ‘single-unit’ or ‘multi-unit’ activity manually.

The following criteria were considered to classify clusters:

- Waveform amplitude and shape: a cluster reflecting a single-unit should have an amplitude considerably larger than the baseline noise and a triphasic waveform.
- Autocorrelogram (ACG): there should be evidence for a refractory period in the cluster’s ACG of spike times. Ideally, for a cluster to be a single-unit, the ACG should be 0 at the centre of the x-axis corresponding to an extremely low or non existent correlation at lag 0.

In addition, we merged of several clusters into one cluster based on similar criteria:

- Crosscorrelogram (CCG): Only if the CCG showed evidence of a refractory period, then two clusters could be considered as containing the spike times of the same unit.
- Waveform shape: the waveforms of the clusters should be similar and highly correlated if they contain spike times of the same unit.
- Location of the clusters in the probe: the spike times should come from close spatial localizations in the probe if they come from the same unit.

2.1.6 Histology and recovery of probe trajectories

After the recordings, we injected mice with an overdose of ketamine and medetomidine. Mice were perfused with saline and then with 4% paraformaldehyde. We cut brains in 100-μm coronal sections on a vibrotome and stained brain slices with 49,6-diamidino-2-phenylindole (DAPI; Vectashield) to visualize nuclei. We performed wide-field microscopy on the slides (Axio Scan Z.1) to verify the location of the probe insertions and register them to the brain atlas (see below). Probe trajectories were reconstructed from histology using publicly available custom code [77]. Probe trajectories were estimated from wide-field images by manually identifying the probe in the image and transforming the location into Common Coordinate Frameworks (CCF) coordinates [78]. The CCF coordinates were then mapped to brain regions in the Allen Brain Atlas.

2.1.7 Data preprocessing

Behavior files consist of experimental and behavioral data on the current session. Neural activity for each session is contained in single unit data files. Only units qualified as 'single-units' were selected for the single-neuron analysis. We also only kept neurons whose presence ratio in the session recording is higher than 0.7. Presence ratio is the fraction of time during a session in which a unit is spiking [79]. We included all neurons available for the population analysis.

Data format consists of the time points, with respect to the start of the recording, at which event of interest happen (odor on, reward on, licking, neuron spiking). Recording licking count is binary, 1 to indicate licking in the bin, 0 to indicate absence of lick, with an original 33Hz sampling rate (0.03s/bin). We re-sampled the licking to our sampling rate, i.e. 10Hz (0.1s/bin).

2.2 Data analysis

The code for the analysis can be found at the following GitHub repository:
https://github.com/CeliaBenquet/belief-states.

Unless indicated, the neural data analyzed in the project come from the same recording. Neurons recorded in the OFC are from the lateral OFC, i.e. ORBl 2/3, 4, 5 and 6a. Neurons in the secondary motor cortex (M2) corresponded to the MOs 1, 2/3, 5 and 6a. Neurons from the primary motor cortex (M1) were from a different animal and are mapped in the MOp 1, 2/3, 5, 6a and 6b. Acronyms taken from the Allen Brain Atlas [78].

2.2.1 Peak activation maps

Neurons were sorted on the timing of their trial-averaged maximum spike count, separately for each odor type. Sorting was performed on a window of time from 1s prior to odor onset to 1s after reward onset to compute the peak activations, on half of the trials. Then, we displayed neural activity for a window of 2s prior to odor onset and 4s after reward onset (4s being the minimum early ITI before the following trial starts) on the other half.

Neurons were ranked based on the index of the time bin presenting their maximum activity, from earlier to later peaks. Neurons in the validation set were sorted using the ordering found on the training set. Resulting activation map was z-score scaled per neuron across trials ($X_{norm} = (X - \mu)/\sigma$) so that the activity pattern is not affected by noise.

2.2.2 Permutation test

For a given cue, we randomly separated available trials into two equally-sized sub-groups. We found the peak response times of each neuron on the ISI (between odor onset and earlier reward) for each subset. We ranked neurons from earlier to later peak response in each. We found the Spearman's rank correlation between both orderings, using the SciPy Python library [80].

To get the distribution, we reiterated the following process for 1000 random draws of trials from the recording. To determine if the correlation is significant, we performed a shuffle test. We computed the probability that the true distribution of the Spearman' rank coefficient is higher than a null distribution. For that, we computed the null Spearman's rank correlation distribution obtained by chance, by finding the Spearman's rank correlation between the first subset ordering and the second subset ordering, shuffled, for all epochs.

2.2.3 Regression matrix and linear regression encoder

We designed a multiple linear regression model between the task variables, as input variables and the neurons firing count as output variable. We computed the least-square solution w that solves the equation $x \times w = y$ with x, the regression matrix containing the task variables, y, the neurons' firing rates and w, the weights associated to each task variables. We used the NumPy.linalg Python library for the implementation [81].

Regression matrix

We investigated the influence of four computational variables on the neural data, i.e. odor, delay after odor, reward onset, licking. Each variable was convolved to bases to construct the regressors used to form the regression matrix. The regressors are constructed so that they represent time-lagged versions of each variable. It allows them to account for the neural signal response to particular feature patterns in time. Odor onset and delay after odor are regrouped into the same variable, *the ISI variable*, starting at cue onset and capturing neural activity between cue onset and reward onset. Earlier regressors (4 first regressors; 0.8s from cue onset) represent the odor variable. Later regressors (10 last regressors; from 0.8s after cue onset to reward onset) represent the delay variables. The reward variable is tuned on reward onset and spans for 3s seconds to capture neurons activated upon reward. The licking variable spans for 1s after lick onset, capturing neurons tuned to licking. A nuisance variable spanning the entirety of the recording was added to remove the effect of drift from the other variables. Parameters for those variables and the geometry of the bases we designed for each to construct their regressors are summarized in Table 2.1.

For the ISI variable, we designed the bases as a sequence of sub-states marking post-stimulus time-steps. As time elapses, we considered that the belief state proceeds through the ISI, dividing it

Variable	Event	Basis type	Basis duration	Bases number
ISI	Cue	unit	min 1.2s, max 2.8s	14, cut at reward
Reward	Reward	unit	3s	15
Licking	Licking	cosine	1s	5
Nuisance	Recording onset	cosine	Full recording	5

Table 2.1: Predictor variables for the linear regression analysis regression matrix. The regression matrix consists of all or a reduced combination of the presented explanatory variables, each convolved to their corresponding basis. The ISI variable is tuned on cue onset and separated into earlier and later regressors for odor and delay variables respectively.

into 14 states of 0.2s each, spanning from cue onset to reward (Fig. 1.7) [14]. Similarly, we designed the variable to start at cue onset and span on the entirety of the ISI, until reward happens. Bases are ”unit” vectors, where 2 subsequent bins per basis are set to one, similarly to Figure 1.7. Values after reward are zeroed for all regressors of the variable. The resulting regressors in the matrix form a ramp following each cue and cropped at reward. We also used the same unit bases to construct the reward regressors, with 15 bases (grid-search from 5 to 19 bases was performed to assess performances; Supp. Fig. A.4), spanning time for 3s after reward. We used cosine bases to construct the licking, odor and nuisance regressors.

Data splitting

Data splitting was performed across trial types. Splitting was performed on full trials with a 70:30 train-test ratio. Validation set always contained all trial types, even when training was performed only on one type of trial, for instance only odor A and only cue C models discussed when investigating odor specificity.

Quantifying models predictions

Models were fitted for each neuron. We studied the influence of each variable by observing the fitted weights and predictions. The fitted models were used to predict spike traces $\hat{y}$ on validation data. Models performances were evaluated by computing the adjusted R^2-score of these predictions:

$$adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{2.1}$$

where the R^2-score, between the true data y and the predicted data $\hat{y}$ was computed using the sklearn Python package [82]. The R^2-score is the rate of the variation in the data that is predictable from the (independent) variables of the model used to obtain the predicted data.

F-test

An F-test between full and reduced models for each of the variables presented in Table 2.1 was performed. In other words, we tested the significance of a variable in explaining the data by testing whether a model without the said-variable could predict spike counts statistically as good as the model with all the explanatory variables. If it could perform as good as the full model, then the variable was qualified to be non-significantly encoding the neuron, with respect to the full model. We stated the null hypothesis H_0 as *the reduced model is able to explain as much variance in the data as the full model and the removed predictor is non-significant* while the alternative hypothesis H_A is *the predictor is significant as the reduced model is not able to explain as much variance in the data as the full model without it*. We trained each of the models and predicted spike trace from the fitted weights. Then we computed the F-value between each of the full-reduced model combinations:

$$F^* = \frac{(SSE(R) - SSE(F))/P}{SSE(F)/(N - P - 1)} \tag{2.2}$$

with R the reduced model, F the full model, $SSE(m) = \sum_i (y_i - \hat{y}_i)^2$ the sum of squared errors using model m to get predictions $\hat{Y}$, P the number of extra encoding variables in the full model compared to the reduced model, i.e. here always 1, and N the number of observations. We used the SciPy.stats Python library to find the F-cumulative distribution function (cdf) associated to the F-value we found. We deduced the p-value $P = 1 - cdf(F^*, P, N - P - 1)$ [80]. If the p-value was small enough (lower than $\alpha = 0.005$) then there was statistical evidence to reject the null hypothesis. It meant that the variable was statistically significant with respect to explaining neural activity as good as the full model.

The F-test only indicated if the variable is significant compared to the full model. Hence, we made the strong assumption that the full model is, itself, significantly explaining the data. By using that F-test only, a neuron could be found significantly encoding a particular variable, while the full model was actually performing poorly in explaining the neural activity. To leverage this assumption, we re-defined a neuron as significantly encoding a given variable if the reduced model was rejected on the F-test but also, the adjusted R^2-square score on the full model is sufficiently high (threshold arbitrarily set to 0.05). We summed the neurons found significant for each variable to get the fraction of neurons significantly encoding it in the studied region. We performed such a test for all full/reduced models combinations and all regions.

General Linear Model (GLM)

We also performed a GLM analysis [83]. We preferred the linear regression model. It had a lower computational cost, as it did not require to tune hyperparameters. Moreover, performances were found qualitatively similar (Supp. Fig. A.6). We implemented an elastic-net Poisson GLM [84]. Elastic-net is a regularization method that linearly combines L1 and L2 penalties, overcoming limitations of each penalty, while Poisson GLM is able to model count data such as neuron spike rate, following the Poisson distribution. We used the pyglmnet Python package [85] for the implementation. Maximum likelihood estimation was used to estimate the regression coefficients. Similarly to the linear regressor model, we fitted one model per neuron. The GLM hyperparameters were tuned using cross-validation on the training set.

2.2.4 States decoder

A multi-class linear classifier was trained on spike counts ($R^{T \times D}$, with T the number of bins and D the number of neurons) to estimate the true experimental state (R^{T}) at every time step. We used the sklearn.linear_model Python library [82]. Data consisted in the z-score standardized binned spiking count (sampling rate of 10Hz). As the animal spent most of a recording in the ITI, the states were first manually re-balanced so that we had the same number of ITI samples as the number of available samples in the biggest ISI states (states in early ISI). We split the sampled bins randomly across trials on bins of same time steps with a splitting ratio of 80:20 train/test. Validation data correspond to a snippet of 6 trials (20s; 200 bins) on a time window of time from 0.2s prior to odor onset to 0.2s after reward onset.

We developed two decoders: a macro-state and a micro-state decoder. In the first one, states were simply the ITI and the ISI. The second one incorporates temporal sub-states, similarly to a CSC mode, and inspired by Figure 1.7 [14]. Those sub-states were designed so that they parse the ISI into 14 micro-states while ITI is the 15th state. Performances for the first decoder were evaluated using an area under the receiver operating characteristic curve (AUROC) metric, which is a measure of the ability of a classifier to distinguish between two classes, and we plotted the predicted states against true states. For the second decoder, we computed the confusion matrix, whose rows we divided by the number of occurrences in each classes. It provides us with the true positive rate (or recall) on the diagonal, while the sum off-diagonal values per row gives us the false negative rate (or miss rate). We also computed the predicted states against true states.

2.2.5 Dynamical system model of neural data

We used a recurrent switching linear dynamical systems (rSLDS) model [86] to approximates the complex non-linear dynamical system formed by the neural population activity. This model decomposes complex non-linear systems into a set of linear dynamical systems. The ssm Python library (Linderman lab) provided the framework to implement such a model [87].

The model relates three sets of variables: a set of discrete states z, a set of continuous latent factors x that captures the low-dimensionality of the neural activity, and the activity of recorded neurons y. It also allows for external inputs u for which we used the presence of cue and reward in current time bin, separating cues to one row each if recording includes multiple trial types. At each time step t, the model associates a discrete state z depending recurrently on the continuous latent factors and external inputs so that $z_t | u_t, z_{t-1}, x_{t-1}$ is sampled from a probability driven by a logistic regression on continuous state:

$$p(z_t = i \mid z_{t-1} = j, x_{t-1}) \propto \exp\left(R x_{t-1} + W u_{t-1} + r\right) \tag{2.3}$$

with R, W and r that parameterize a map from the previous discrete state, continuous state and external inputs to a distribution over the current discrete state, using a softmax link function. The discrete state z_t specifies a set of linear dynamical systems and determines which one to use to generate the continuous latent factors so that $x_t \in R^M$ is defined as:

$$x_t = A_{z_t} x_{t-1} + V_{z_t} u_t + b_{z_t} \tag{2.4}$$

where $A_{z_t} \in R^{D \times D}$ is a dynamics matrix and $b_{z_t} \in R^D$ is a bias vector with D the dimensionality of the latent space. Finally, the activity of the recorded neurons is modelled as a linear noisy Gaussian observation $y_t \in R^N$ with N the number of recorded neurons: $y_t = Cx_t + d$, with $C \in R^{N \times D}$ and $d \sim \mathcal{N}(O, S)$, a Gaussian random variable.

Overall, the model needs to fit the state transition dynamics, the linear dynamical system matrices and the neuron-specific emission parameters, summarized as $\delta z_t = \{A_{z_t}, V_{z_t}, b_{z_t}, C, d, R, W, r\}$. These parameters are tuned optimizing a maximum likelihood that is using a combination of Laplace approximation and variational inference methods. The model performance is reported as the *evidence lower bound (ELBO)* (evidence of lower bound of the log-likelihood; higher is better), equivalent to the Kullback-Leibler divergence between the estimate and true posterior $KL(q(x, z; \phi)||p(x, z|y; \theta))$ using 5-fold cross-validation [88].

Neural data was always first z-score standardized over the whole recording for each neuron. We provided neural activity and inputs per trial to the rSLDS, of size ($T \times B \times N$) and ($T \times B \times M$) respectively, with T the number of trials, B the number of bins per trial, N the number of neurons and M the number of input dimensions. We performed grid-search on K, the number of discrete state, from 1 to 6, and kept the K for which the average ELBO over 5 fittings was the highest (Supp. Table A.1). We kept $K = 4$. We selected 2 as the number of latent variables, as we wanted to be able to display the resulting latent space in two dimensions.

rSLDS on RNN hidden activations

We made use of the noisy hidden activations generated by the pre-trained GRU network of three hidden units, when performing T trials of B bins per trial,sampling rate 0.2s, on odor A only. To make the activation layers noisy, noise was added at weights update during training. To generate the neural activity for N neurons over $T \times B$ bins, the following generative model was applied: $spikes = A \cdot activations + b$. $spikes \in R^{N \times (T \times B)}$ is the neural activity over the session, $A \in R^{N \times 2}$ is a matrix randomly generated from a normal distribution $\mathcal{N}(0, 1)$ corresponding to the factors weighting for each individual neuron, $activations \in R^{2 \times (T \times B)}$ is the matrix containing the hidden activations and $b \in R^{N \times (T \times B)}$ is the noise, that we fix to be null, in order to get neural data as close as possible from the activation. Inputs $u \in R^{2 \times (T \times B)}$ to the rSLDS were the same as the ones provided to the RNN. They consist in the observable environment, meaning whether odor, reward or no event was happening at each time point, with one channel for cue and one channel for reward. Data and inputs were grouped into trials to be fed to the model, their final size being respectively ($T \times B \times N$) and ($T \times B \times 2$).

The model performances were measured using the ELBO metric. We trained the model for a few iterations and chose the best one, as initialization of the discrete states and continuous latent factors varies from one iteration to the other.

Removing drift from the population activity

We studied the evolution of the trial-averaged activity throughout the session recording for all neurons, using dimensionality-reduction (PCA). We observed that there was a global drift through the recording, mostly explained by the two first principal components (Supp. Fig. A.14).

Considering $y \in R^{(T \times B) \times N}$, the neural data, with T, the number of trials of B bins each and N, the number of neurons, we removed the global drift from this z-score standardized neural data set by applying the following transformation:

$$y_{no\_drift} = (y - y_{psth}) \times (V^T_{no\_drift} V_{no\_drift}) + y_{psth} \tag{2.5}$$

with $y_{psth} \in R^{B \times N}$ the peristimulus time histogram (PSTH) for all neurons over the whole recording and V_{no_drift} the principal directions of the data without the two first principal components. We obtained V_{no_drift} by performing a SVD decomposition on the trial-averaged activity for trial and each neuron, studied previously using PCA, and looking at the resulting V matrix columns. In other words, we projected the inter-bins activity variation ($y - y_{psth}$) to a dimension without the drift across trials in time ($V^T_{no\_drift} V_{no\_drift}$) to which we added the mean activity again.

rSLDS on neural population activity

The quality of individual neuron recordings was crucial to the single-neuron analysis. That is why we previously removed units that were classified as ’multi-unit’ from the data and kept only the ’single’ units with a presence ratio higher than 0.7 (see Sec. 2.1.7). Here, however, we were only interested in the behavior of the population and we consequently included all neurons available. Population consists of 77 neuron units. We applied drift removal to this set of neural data and fitted an rSLDS model, similarly to the previous section on the GRU activations. As the dataset includes all four odors, the input matrix is constituted of five channels, one for each odor plus one for rewards. Again, we applied grid-search to determine the best discrete states number, which was found to be 4, and selected 2 as the number of latent spaces.

Chapter 3

Results

To study the dynamics of the belief state, when facing a decision during a reward learning task with state uncertainty, we trained mice on the variable reward delay task. It consists of two olfactory Pavlovian conditioning tasks with four cue types. In task 1, reward-predicting odors A, B and C forecasted reward delivery in 100% of trials. In task 2, odors A, B and C forecasted reward delivery in 90% of trials. In both tasks, the ISI following odor A was drawn from a discretized Gaussian distribution (2.0 ± 0.5s, mean ± s.d.) defined over nine time points ranging from 1.2s to 2.8s. Trials for odors B and C had constant ISIs of 1.2s and 2.8s, respectively. Odor D trials were never rewarded.

In that paradigm, states could be defined as being in the ITI or ISI. In task 1, cue onset and reward onset unambiguously signaled the transition between the states (Fig. 2.1.b.). Consequently, a belief state representation was identical to a CSC state representation. In task 2, reward omission caused the ITI state to self-transition. It resulted in both ITI-ISI and ITI-ITI generated the same observations, making the states hidden (Fig. 2.1.e.). Based on previous work on the task, we assumed that the two macro-states comprise temporal sub-states, or micro-states, similarly to the CSC model [14]. Analogously, we parsed the ISI with 14 sub-states of 0.2s each (sampling rate is 10Hz, 2 bins per state; Fig. 1.7).

Using Neuropixels 1.0 [60], the activity of individual neurons was recorded in the PFC, and more specifically the OFC, M1, M2 and PIR cortex, one of the olfactory cortices [89]. Animal training was assessed based on anticipatory licking, that consistently increased when getting closer to reward (Fig 3.1.a.).

3.1 Single-neuron activity shows specific task-variables tuning in the OFC

Before performing pooled population-level analyses, we first studied the tuning of individual neurons in the OFC. We compared it to M1, M2 and PIR cortex. We first analyzed trial-averaged activity of single neurons in the OFC (Fig. 3.1), PIR cortex, M1 and M2 (Supp. Fig. A.1).

In the OFC, we found numerous patterns of activity from one neuron to the other. We could

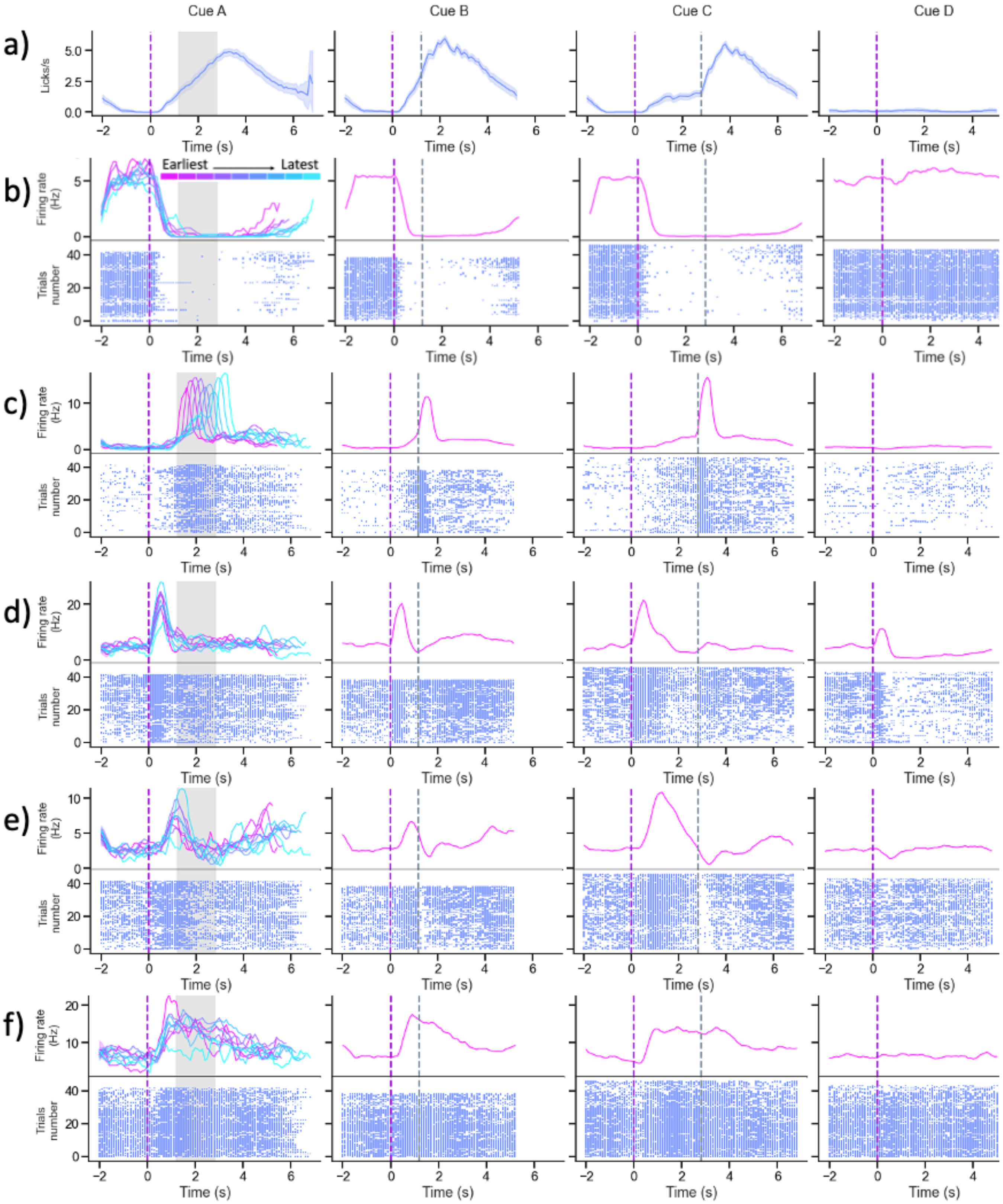


Figure 3.1: Licking behavior and example OFC neurons encoding task-relevant variables of the environment. Trial types are separated. Curves correspond to trial-averaged lick count and firing rates, smoothed with a 1-s window 0.1-step moving average. Time zero corresponds to odor onset and is indicated by a purple dashed line. Reward onset is indicated by either a grey area or dashed line. **a.** Lick count per second. Shaded area corresponds to the s.e.m. **b-f** Example individual neurons showing a diversity of activity patterns in reaction to the task components, i.e. cue and reward. Upper plot is the trial-averaged firing rate (Hz) per cue. Lower plot is the spike raster per trial type, with trials course from bottom to top. For cue A, the trial-averaged firing rate are distinctly calculated for each reward timing and displayed from pink to blue with pink being the earliest possible reward (3.2s) and light blue the latest possible reward (4.8s). **b.** Example background neuron, inhibited in rewarded isi, with activation coming back at fixed time point, differing for each reward timing. **c.** Example reward-tuned neuron, bursting at reward timing. We see a clear ordering of reward timings in cue A trials. We do not see specific activity for unrewarded trials (odor D). **d.** Example odor-tuned neuron, activated at odor onset, even for unrewarded trials (odor D). Magnitude of the activation varies with the expected reward delay. For rewarded trials, the later the expected delay is the larger the activity is. **e.** Example short delay-tuned neuron, activated at the same time step after cue onset regardless of the reward timing, and inhibited at reward. We do not see such a pattern for odor D trials. **f.** Example long delay-tuned neuron, activated at the same time step regardless of the reward timing and staying activated until a fixed time after reward before going back to background activity. No such pattern of activity is visible for odor D trials.

qualitatively classify the neurons into sub-groups. We distinguished four main sub-groups of pattern of activity: odor-tuned, reward-tuned, delay-tuned and background-tuned neurons. Background-tuned neurons (Fig. 3.1.b.) were inhibited at rewarded cue onset and reactivated after reward onset. The reward-tuned neuron showed (Fig. 3.1.c.) a typical burst of activity right after reward onset. We also observed neurons tuned to reward that rather gradually decayed back to background activity and some that got inhibited at reward. Odor-tuned neurons (Fig. 3.1.d.) got activated upon odor onset, with the activity magnitude that varied depending on the trial type and neuron. Delay-tuned neurons showed specific activity in the ISI for rewarded odors but *not* for the unrewarded odor. We observed neurons bursting rapidly (Fig. 3.1.d.) at a specific post-cue onset time step, consistently across rewarded trials and decaying either rapidly or gradually until reward onset. We also noticed neurons whose activity was sustained at a high firing rate plateau before decaying once reward was delivered (Fig. 3.1.f.).

We consequently observed the trial-averaged activity across all neurons. We constructed activity maps, in which neurons were ordered by the timing of their peak responses (Fig. 3.2). We used half of the trials to sort neurons based on the timing of their peak response. Then, we plotted the z-score standardized spike count for the other half, using the same neuron sorting. Of interest, the OFC neurons in the ISI activated in sequence, following the cue onset and spanning the entirety of the ISI (Fig. 3.2.a). Bursting neurons got activated in the early ISI, gradually replaced by the sub-population of neurons showing sustained activity until reward onset. There was no such sequential activity of the neurons' peaks activation in unrewarded trials (odor D). This indicates that it might span the ISI to keep track of time elapsed before reward delivery. It makes it a key-element to represent states, and remembering CSC or even microstimuli features representation.

Such a sequential activity was visible in M2 (Fig. 3.2.b.) while not in the PIR cortex (Fig. 3.2.c.) and M1 (Fig. 3.2.d.).

We further quantified the sequential activity that we observed. Based on our observations, neurons that are part of the sequence should consistently peak at their own time step after cue onset. Hence, in a region where sequential activity is present, the sorting we obtain from one randomly-drawn set of trials to the other should not vary much. Spearman's rank correlation coefficient is a measure of strength and direction of association between two variables. To quantify sequential encoding in the ISI, we calculated the Spearman's rank correlation coefficient on the neurons orderings obtained on two subsets of data. We re-itered for 1000 random-drawings to obtain the Spearman's rank correlation distribution. Then, we compared the true distribution obtained to a null distribution (ordering on the first subset against randomly shuffled ordering on the second). As we observed sequential activity exclusively in the ISI, we bounded the sorting to time steps in the ISI only. To exclude a contribution of reward-tuned neurons, we also only used time periods before the *earliest* reward for all odor-A trials. Consistently with our observations, neurons in the OFC present a significantly similar ordering across trials ($P < 0.001$) for all rewarded odors.

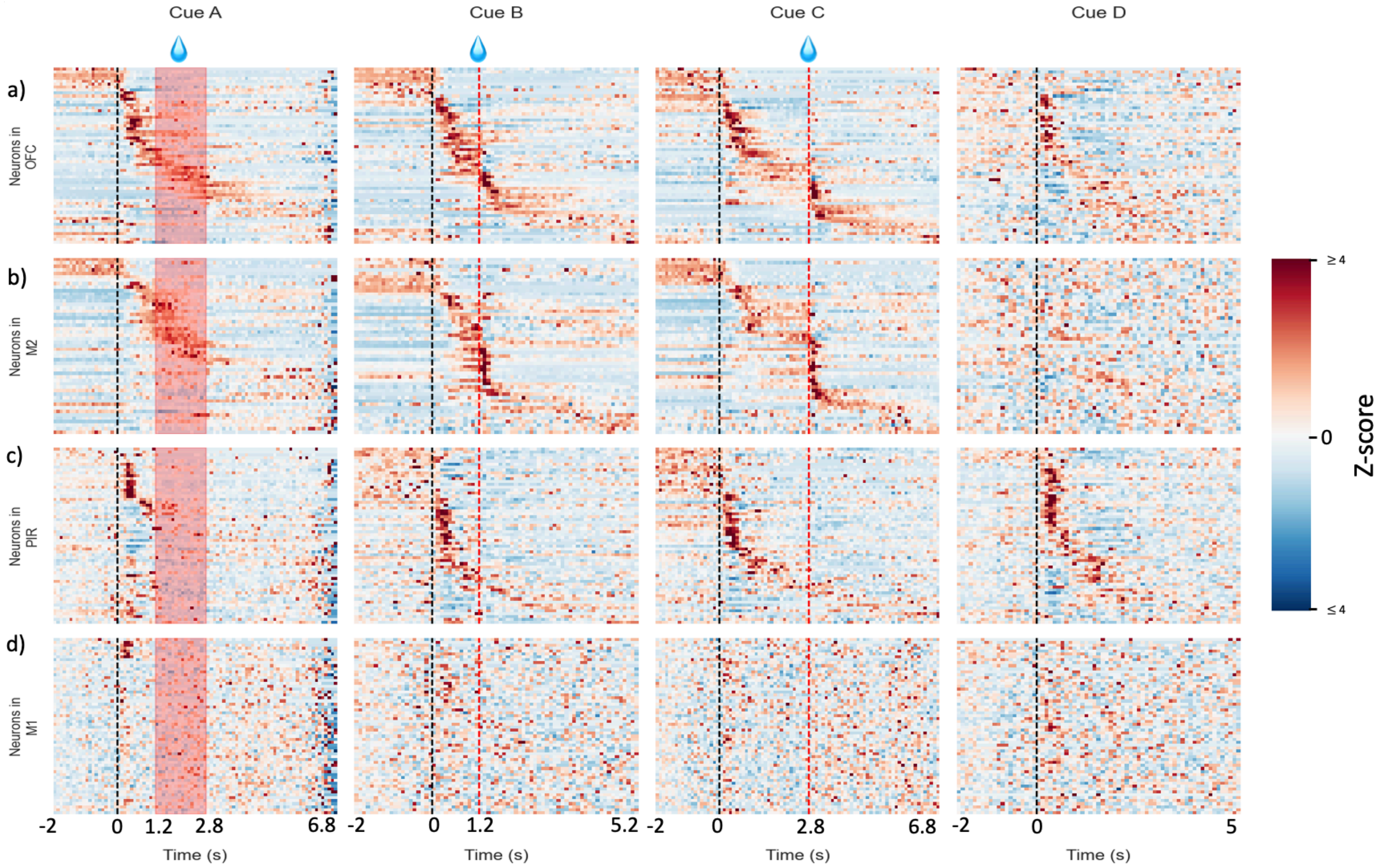


Figure 3.2: Activity maps with neurons ordered per peak response. For each brain region (rows), trial types are separated (columns) and the trial-averaged activity per neuron was computed. Time zero corresponds to odor onset, indicated by a black dashed line. The reward onset is indicated by either a red shaded area or dashed line. For each brain region, neurons are sorted based on the timing of their peak responses from 1s before cue onset to 2s after reward onset. Sorting was performed on half of the trials for each trial type. We display averaged activity on the other half on a larger time window from 2s before cue onset to 4s after reward onset. Neurons are sorted using the ordering found on training activity, from top to bottom. **Supplementary Figure A.2** displays the activity map on the training data. For each region, we display neuron activity for respectively $n = 59, 51, 60, 60$ neurons for **a.** OFC; **b.** M2; **c.** PIR cortex; and **d.** M1. We confirm the sequential activity in the OFC and M1 by displaying neural activity in those regions for a different animal in **Supplementary Figure A.3** Color code on the z-score values is bounded from -4 to 4, with values above 4 or below -4 thresholded to those bounds.

3.2 OFC encode variables that are essential to belief state representation

We next analyzed which information individual neurons encoded using a linear regression model. We also constructed a commonly-used approach with a GLM. Here, we presented results obtained with the linear regressions for simplicity and because performances on both models were found qualitatively similar (Supp. Fig. A.6). Briefly, we regressed the spike count of each neuron by a set of variables, including external cues (odor onset and reward timing), conditioned behavioral responses (lick count) and intrinsic baseline drift across time. We included time-lagged versions of each variable to account for the neural signal response to particular feature patterns in time. Those shifted versions were referred to as the set of regressors for each variable (Table 2.1 for detailed structure of the regressors).

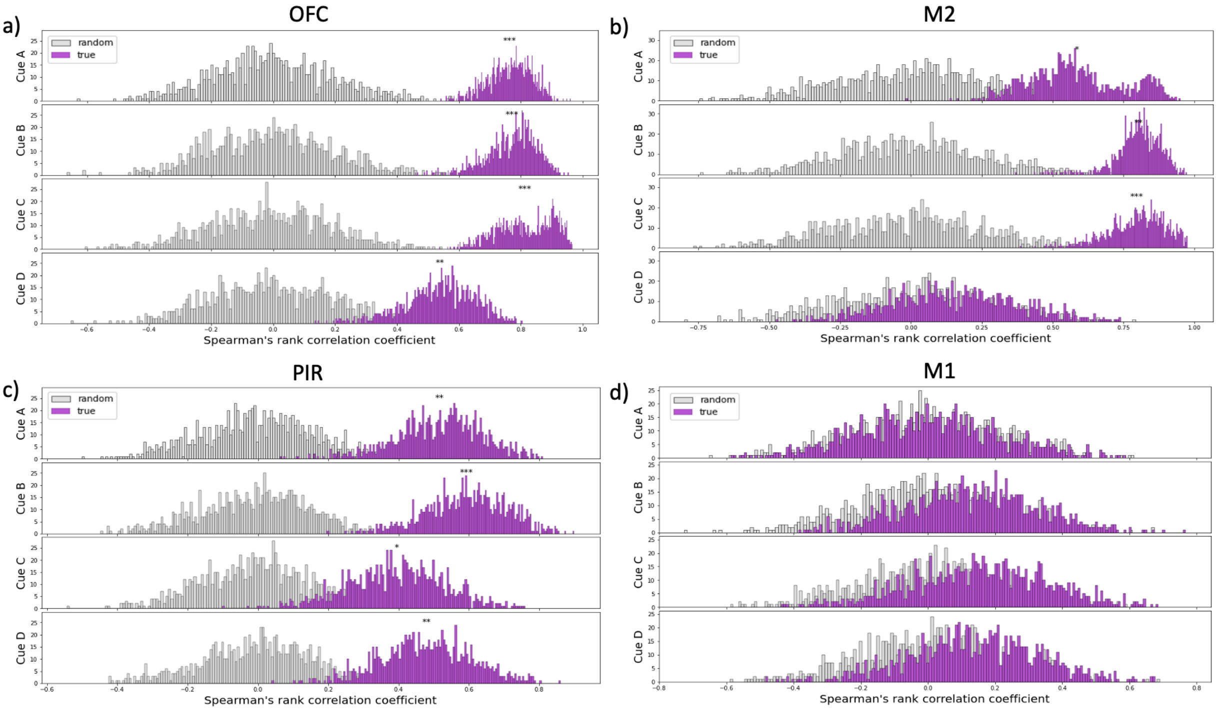


Figure 3.3: Spearman's rank correlation coefficient distribution on the neuron peak response in the ISI. We computed the Spearman's rank correlation between two randomly sampled trials subgroups and applied a permutation test by comparing the true distribution to a randomized distribution per cue type. We consider time points only in the ISI meaning starting at cue onset and ending one bin before (earlier) reward timing. Distribution was computed from 1000 resamplings, for neuron population in **a.** OFC; **b.** motor areas (MO)s (M2); **c.** PIR cortex; **d.** MOp (M1). One-tailed shuffle test, * = P < 0.05, ** = P < 0.01, *** = P < 0.001.

Variables considered in the linear regression were odor onset, reward onset and licking (Fig. 3.4). We designed one set of regressors per cue type and corresponding reward type. The variable capturing activity following cue onset, that we referred to as the *ISI* variable, was further considered as accounting for two types of neuronal sub-population, i.e. the neurons tuned to odor and those tuned to a specific delay after odor onset. For that, we considered early regressors and late regressors of the odor variable (respectively 4 regressors, 0.8s and 10 regressors, 2s) respectively corresponding to an odor variable and a delay variable.

We designed the bases used to construct the ISI regressors so that they represent the sequence of sub-states in the ISI, marking the post-cue time steps (Fig. 1.7). We aimed at capturing the sequential neural activity observed in the previous section (Fig. 3.1). Thus, we tiled the odor-stimuli vector with unit bases so that each one picked out the neural activity structure at the specific moment it was reacting to, post-stimulus. The resulting set of regressors *equally* parsed time bins, forming a ramp along the ISI (Fig. 3.4, top). We constructed the reward regressors similarly, so that the regressors formed a ramp after reward, in the early ITI.

The regression model was used to find the fraction of neurons encoding for each of the designed variables in each region (Fig. 3.5). For that, we designed reduced models, in which we removed one of the variables from the regression matrix. Then, we compared performances to the full model, using a F-test to find the fraction of neurons that are significantly encoding each of the variables.

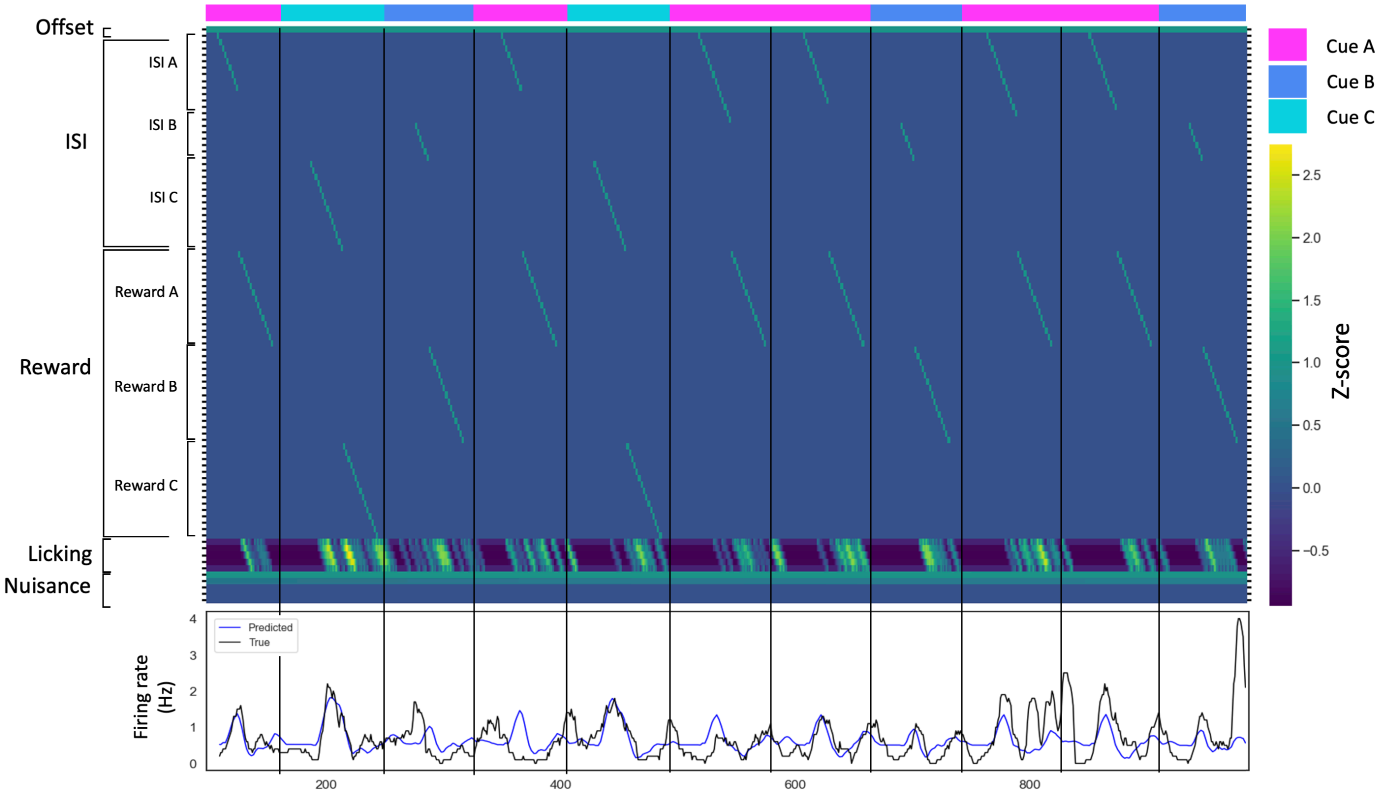


Figure 3.4: Example snippet of the regression matrix containing all variables, corresponding firing rate and prediction across 11 trials for an example neuron (short-delay-tuned; **Figure 3.1.e.**). The trial type is indicated above the regression matrix. The matrix was z-score standardized and we added an extra regressor for the offset variable. Variables are depicted on the left. Bottom plot displays the firing rate (black) and the corresponding prediction (blue) for the example neuron.

The incentive was that, if the reduced model performed significantly worst than the full model and the full model was performing well, then the removed variable was significantly important to the model performances and so the neuron was encoding the information captured by this variable. Overall, in OFC, a majority of neurons were found to encode each of the variables we tested, and that, significantly more than in all the other regions. Consistent with the individual neurons analysis, neurons are encoding for odor and delay. This analysis also confirmed observations on individual neurons described above: M2 also encodes variables important to the task and PIR cortex mostly encodes odor, while M1 is not tuned to the tested variables.

Hence, the diversity of task-relevant activity patterns, from one neuron to the other in the OFC, reinforced with the observation on our linear regression model, provided strong indication that this brain region is representing task-relevant variables. It was also the only region studied to show both odor-tuned activity and delay-tuned neurons. In fact, those independent task-relevant variables are also relevant to define the state of the task. In task 1 (100% of rewards delivered in rewarded trials), the animal could fully deduce whether and, if so, when, it would receive a reward from odor type while reward delivery indicated the transition between the trials. Thus individual neurons are encoding state-relevant variables. Taking them at the population-level, they could represent the full belief state.

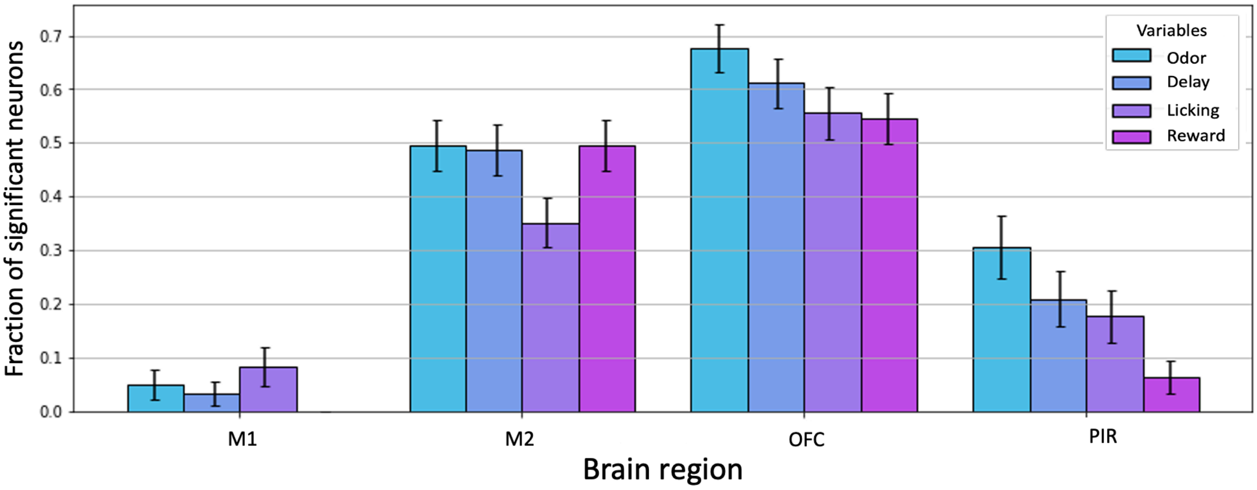


Figure 3.5: Fraction of neurons significantly encoding each variable (see Sec. 2.2.3) listed in the legend for brain region in the horizontal axis. A neuron was defined as significant if (1) its adjusted R^2-square score was good enough (threshold was set to 0.05; see Supp. Fig. A.5) and (2) the reduced model was rejected when performing an F-test between full model and reduced model in which the variable was removed. Error bar is the standard error obtained from the binomial mean variance: $se = \sqrt{\frac{p(1-p)}{n}}$ with p, the rate of significant neurons and n, the total number of neurons for each group.

3.3 OFC population activity is sufficient to predict states

Based on observations at the single-neuron level, we proposed that a full state representation, if it is indeed encoded in the OFC, is encoded considering the whole neural population. Next, we built up on the idea that OFC might encode micro-states, while mPFC would rather encode a form of macro-state by computing the probability distribution over information from the OFC [16]. Considering the belief state representation in which global states comprised temporal sub-states (Fig. 1.7; [14]), we asked if we could find correspondences to that model in the OFC neural activity. If we could decode states, both at the global and sub-states level, given neural activity nearly as good as true experimental beliefs, it would be evidence that neural activity at the population level is belief-like and that in that sense, the neuronal population encodes estimates of hidden states.

To that end, we designed two multi-class linear classifiers based on neural activity in the OFC. The first one predicted the global states (either the ISI or ITI), while the second one predicted the sub-states. We referred to the macro-states for the first classifier, while this second one, for which we separated the ISI into 14 micro-states of 0.2s each (2 bins per state) and the ITI as the 15th state (Fig. 1.7), was referred to as the micro-states decoder.

Both macro- and micro-states decoders performed better, using neural activity in the OFC, compared to the other regions (Supp. Fig.A.8; Supp. Fig.A.9; Supp. Fig.A.10). The decoder on M2 activity performed as good as on OFC activity in early ISI but performances decreased more than OFC as states got closer to reward. Performances on PIR cortex activity were good on the first states and rapidly decreased for the second part of the ISI, which was consistent with the region being tuned to odor. The normalized confusion matrices of the micro-states decoders (Fig. 3.6 Top) showed good performances in predicting accurately the states, mostly at the beginning of the

Trial type	OFC	M2	PIR	M1
All	0.97	0.89	0.80	0.61
Cue A	0.97	0.88	0.82	0.65
Cue B	0.98	0.91	0.91	0.73
Cue C	0.98	0.90	0.84	0.64

Table 3.1: AUROC of the macro-state linear classifier per trial type and brain region. Values were consistently higher for classifier based on the OFC neural activity than on the other brain regions.

ISI and for the ITI. For late ISI, the models lost in accuracy even if they were mostly still predicting states adjacent to the true one in that case. This could be a reflection of the presence of delay-tuned neurons showing a sustained activity waiting for reward in late ISI (Fig. 3.1.f.). Predictions on snippets of the session (Fig. 3.6 Middle-Bottom) showed good performances except on the transitions between the states, especially visible in the macro-state decoder plot (Fig. 3.6 Bottom). Those noisy predictions at transition could be the sign of a change in the overall activity from one stable signature population activity for one of the macro-states to another stable signature activity in the other state. This strongly indicated that the underlying activity of the population contained the necessary components to differentiate between both global states and sub-states.

We consequently further analyzed the population activity, applying PCA dimensionality-reduction to the trial-averaged activity, per trial type. Looking at the resulting low-dimensional representations of the neural activity (Fig. 3.7), we distinguished two regions of the space where population activity seemed to stabilize. The first one corresponded to the beginning of the trial, for the whole duration of the late ISI and until cue onset happened. The second one was at the end of the ISI. It was more

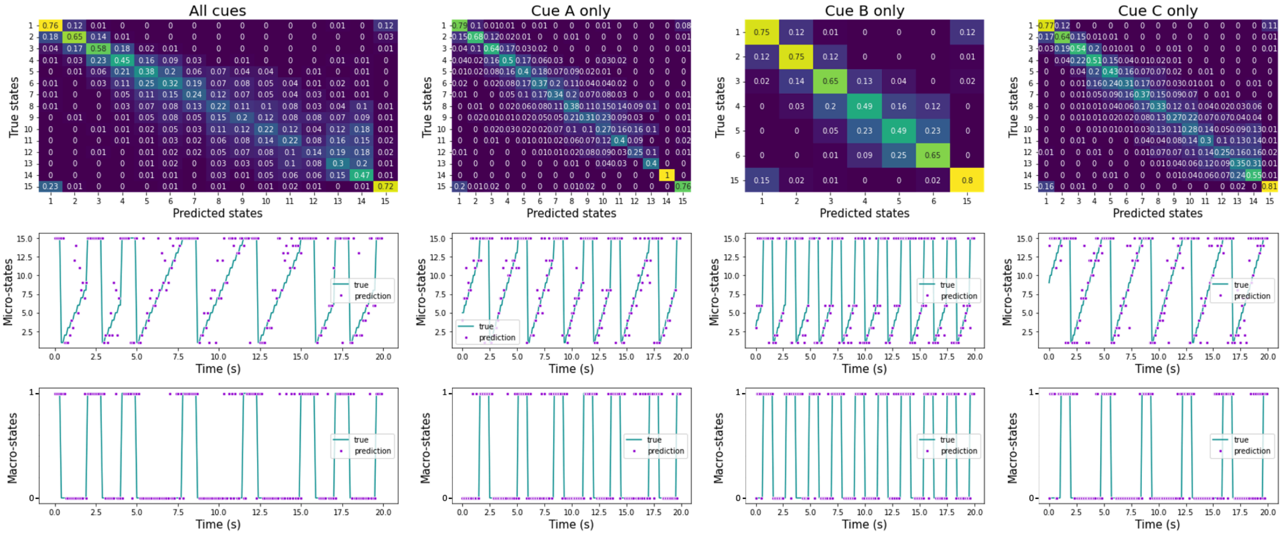


Figure 3.6: State classifier performances from neuron population in the OFC. Per column, we trained the decoders on (1) all rewarded cues, (2) only cue A trials, (3) only cue B trials, and (4) only cue C trials. The same analysis was performed for all groups. **(Top)** Confusion matrix normalized by number of occurrences per state (each cell is divided by the sum on values in its row). Values on the diagonal correspond to the sensitivity ($\frac{TP}{TP+FN}$) for each state. Data evaluated is the testing set. **(Middle)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *micro-state* classifier. **(Bottom)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *macro-state* classifier. As one session mostly consists of ITI, the states were first manually re-balanced in the dataset used for training/testing (same number of ITI samples as the maximum available number of samples in an ISI state. Validation data, used for plots in the middle and bottom rows correspond to 6 trials (20s; 200 bins) for 0.2s prior to odor onset to 0.2s after reward onset.

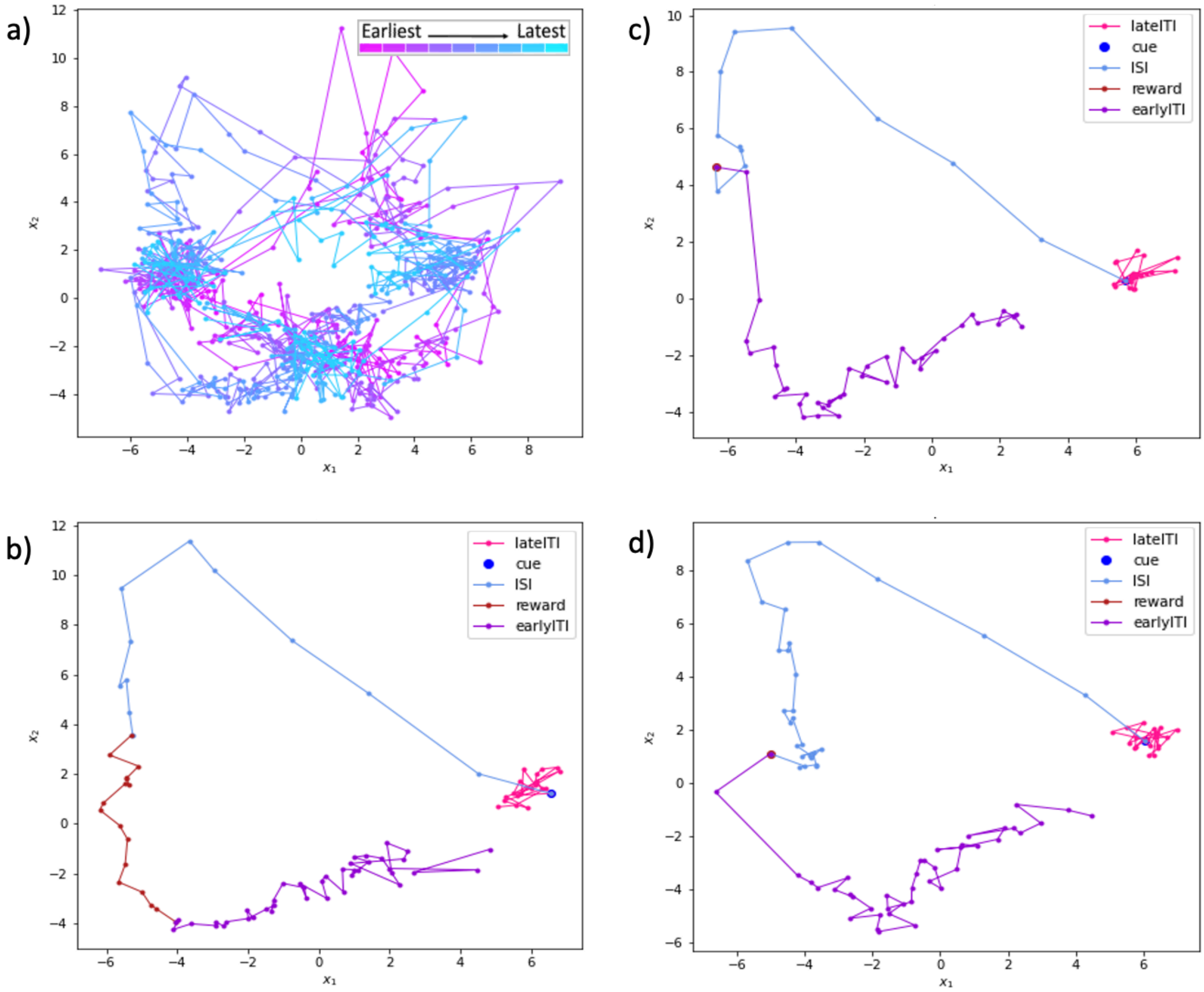


Figure 3.7: Dimensionality-reduction of the trial-averaged neural activity per trial type in the OFC. PCA was applied on trial-averaged neural activity. We displayed the resulting two first principal components. **b-d** Low-dimensional representation of the trial-average neural activity per trial. Late ITI (pre-cue) in pink, cue in dark blue, ISI in blue, reward in red (either point or line), early ITI in violet. For **a.** odor-A trials, computed distinctively per reward type. Purple trajectories represent earlier rewards trials (from 3.2s) and light blue trajectories are later rewards trials (up to 4.2s). Trajectories are noisier than the other plots as the number of samples to fit the PCA were lower; **b.** odor-A trials, reward timings taken all together (from 3.2s to 4.8s); **c.** odor-B trials (reward at 3.2s); **d.** odor-C trials (reward at 4.8s).

visible for longer ISI, i.e. odor-C trials (Fig. 3.7), than for shorter ISI such as odor-B trials (Fig. 3.7). Finally, we observed that the neural activity from cue onset spent an important fraction of the ISI to transition from one region of the plane to the other, before stabilizing. Similarly, in the early ITI the activity goes back gradually to the late ITI region. This reinforced the idea that there were periods of time for change in neural activity at the transitions between the global states.

The decoder performances and the geometrical representation analysis visually corroborated the observations made on the GRU network presented in Fig. 1.8. It learned value estimation by developing two hidden activation layers resembling true beliefs (Fig. 1.8.(a-d)). The trajectory of those hidden activations also showed a consistent spatial separation between bins in the ITI and ISI (Fig. 1.8.(e-f)), similarly to our observations on the low-dimensional space of neural activity in OFC (Fig. 3.7). More importantly, they showed that this geometrical representation is only a description of the underlying topological structure, which consisted of two fixed-points attractors. Those fixed-points were found in place of the ITI and ISI in the plane.

We showed the similarity between the geometrical representation of the neural activity in the OFC and the hidden activations of the GRU network. Moreover, the low-dimensional activity trajectory seemed to stabilize in the two regions of the space corresponding to ITI and ISI, hinting that the dynamics of the activity could also be driven by fixed-points in those regions. However, dimensionality-reduction does not provide the quantification needed to assess the underlying temporal dynamics. Instead, we fitted a linear dynamical system model to discover the dynamical units of the neural activity in a trial.

3.4 Population dynamics in the OFC show two fixed-points corresponding to states

As the dimensionality-reduction analysis indicated two separated regions in the low-dimensional space, we next investigated the presence of fixed-points in the temporal dynamics of the population neural activity, at time of the ISI and ITI. To that extent, we fitted a linear dynamical system model to uncover the dynamical units, i.e. the patterns of ensemble activity, constituting the neural activity observed in our experiment. By analyzing the fit dynamical system, we could infer what type of neural network mechanisms could generate the observed neuronal activity.

The recurrent switching linear dynamical systems (rSLDS) model is a hierarchical, recurrent, and input-driven linear dynamical system model. It approximates complex nonlinear dynamical systems using a composite of linear dynamical systems [86]. It can learn both a flexible non-linear generative model and how to parse the data into coherent discrete units. Thus, once fitted, it provides the dynamic units of the neural activity as well as the relationship of the dynamical switching to environmental stimuli. We used it to uncover the topology of the neural activity space.

We determined the hyperparameters of the model similarly to a previous work from Nair and colleagues (2022) [17]. The number of discrete states was evaluated to maximize the likelihood of the data, using cross-validation. We fixed the number of latent variables to 2, to be able to observe the latent space in a two-dimensional plot. To ensure that the model converged to a global minimum, we reinitialized and retrained the model repeatedly and kept the model with the best loss (highest ELBO).

To comparing the underlying dynamical systems of the GRU network to our empirical data, we first fitted the rSLDS model to the two hidden activation layers of the trained GRU network. For that, we first generated higher-dimensional data, from the two-dimensional internal activity of the GRU network and fitted the model. Note that we had to augment the dimension of our data because the implementation of the rSLDS is such that the dimension of the observations must be strictly higher than the dimension of the latent space. The true activations presented a cyclic temporal dynamic (Fig. 3.8.a.), in which the cue and reward onsets both caused a response in the activity. Stimulus drove the activity from one state to the other. Reward response drove the trajectory to abruptly stop its progression towards the ISI fixed-point in earlier reward delivery timings. In later ones, the trajectory

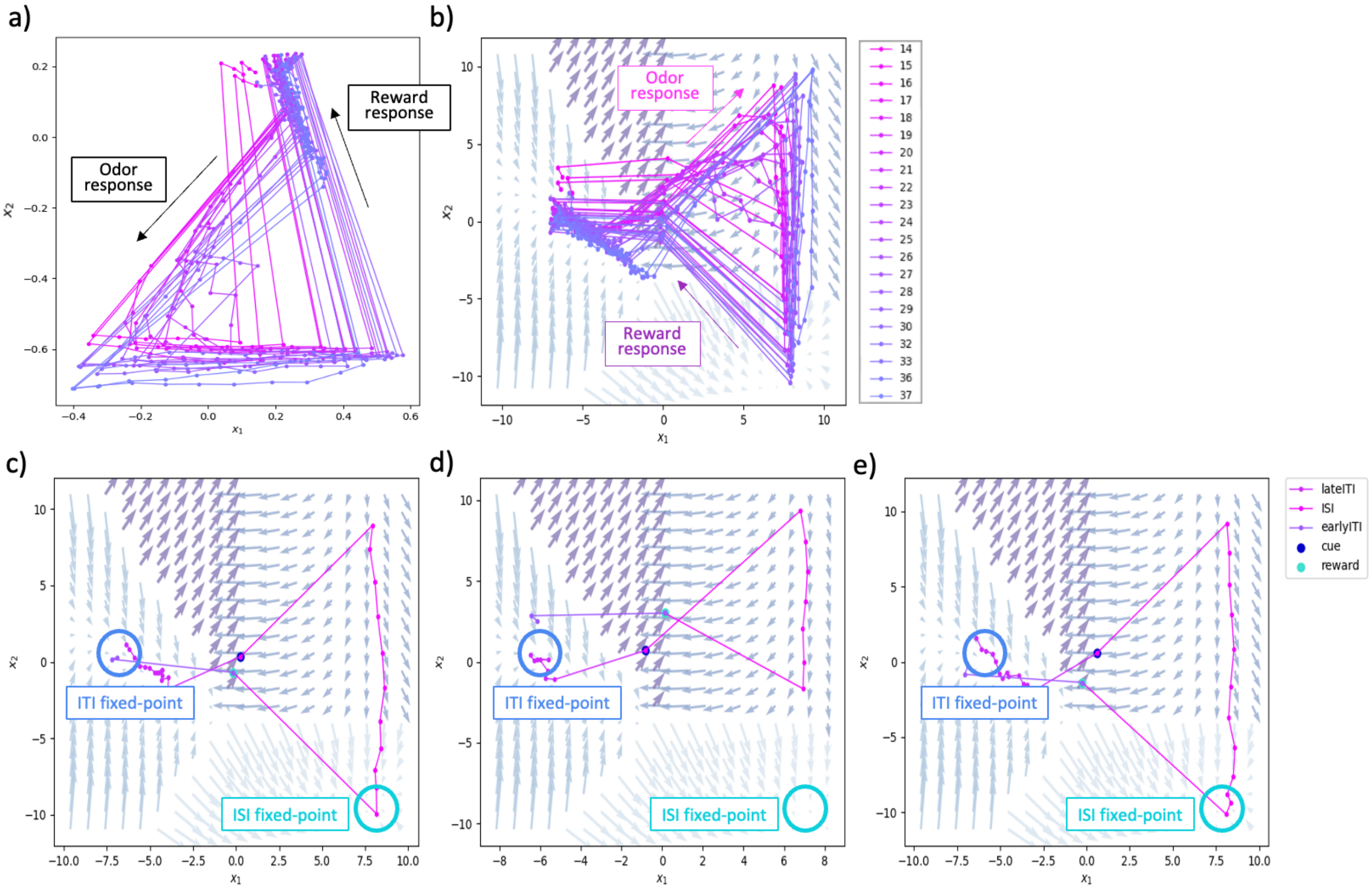


Figure 3.8: GRU activations trajectories and neural activity dynamics. a. Trial-averaged true GRU activations, per reward timing, plotted against each other. We see the two fixed-points right before odor response and right before reward response. Reward timings are distinguishable as early reward responses cut the activity trajectory before reaching the ISI fixed-point, while for later rewards, the activity stays at the fixed-point until reward happens. **b.** Trial-averaged latent trajectory resulting from the high-dimensional data generated from the activations, per reward timing, obtained using an rSLDS model. Flow field correspond to the 4 linear models used to fit the data, each one corresponding to a discrete state. Legend corresponds to the bin number in the trial at which the reward happens, sampling rate is 0.2s. Trials consist of a 2s-late ITI, followed by the ISI, of varying size, from 2.8s (14 bins) to 7.4s (37 bins), randomly chosen, following a Gaussian distribution, followed by a one-bin (0.2s) early ITI. **c-e** Example individual trials, with difference size of ISI, sampled randomly from the recording. Flow field corresponds to the most likely dynamics in the plane, with each color representing one of the 4 linear systems fitted (one per discrete state) and arrows show the direction and rate of change of activity at each point. Fixed-points are identified as the regions of the flow field showing zero-length arrows. ITI and ISI fixed-points are identified in the plan by the respectively blue and turquoise circles. Model showing the best loss was kept.

reached the fixed-point and stabilized there until reward. The corresponding rSLDS latent space trajectories (Fig. 3.8.(b-e)) showed the same characteristics. We could appreciate the difference in the trajectory velocity and length in the ISI, depending on the reward delivery timing (Fig. 3.8.d. for earlier timings; Fig. 3.8.(c,e) for later ones). The underlying flow field also assessed the presence of fixed-point attractors in place of the ITI, before the odor response, and ISI, before the reward response.

Next, we fitted the rSLDS model to the OFC neural activity. The averaged spiking count per trial across the session showed a consistent drift in the population activity as time passed, mostly on the two first principal components (Supp. Fig. A.14). As it could not be explained by the task, we removed the drift from the z-score scaled neural data, by projecting it to a lower dimensional space, from which we removed the 2 first dimensions. We fitted the model for around 100 iterations and selected the best fitted model (highest ELBO; Fig. 3.9). Observations were similar to what we observed in the dynamics of the activations layers of the GRU network.

The latent dynamical space for the best fitted model showed two fixed-points that drove the trajectory, during the ITI in the purple linear space, and ISI in the blue linear space, respectively. In the ISI, we saw a change of velocity in the dynamics of the activity, depending on the estimated time of reward delivery. The trajectories for earlier reward delivery trials were faster to move to the ISI fixed-point than for later reward delivery trials, when responding to odor onset (Fig. 3.9.b., Cue B trials against; Fig. 3.9.b., Cue C trials; trajectory was more or less on the inside of the loop). We also observed the same dynamics as for the GRU activations regarding the ISI fixed-point. For Cue A trials, activity trajectories for each reward delivery timing responded to reward in order, from earlier to later reward delivery timings. Later reward delivery timing trajectories stabilized in the fixed-point. Earlier ones were driven to it but stopped when reward delivery happened (Fig. 3.9.a.; Cue A). Finally, early ITI did not seem to be coming back to the ITI fixed-point region, where the trajectory was stabilized before cue response. We confirmed this observation with Supplementary Figure A.16, in which dynamics were such that the the early ITI is ”recaptured” by the ISI fixed-point. However, dynamics in the the early ITI in Supplementary Figure A.17 did go back to the initial late ITI. Hence, it might have been due to the short period of the early ITI that we provided to the model.

We showed that the neural population activity in the OFC displayed an attractor-like dynamic, constituted of two fixed-point in place of the two macro-states in our task. Moreover, the neural activity dynamics were similar to those found on the hidden activation of GRU network that were found to be belief-like. Overall, we have strong evidence that the OFC is encoding state representation.

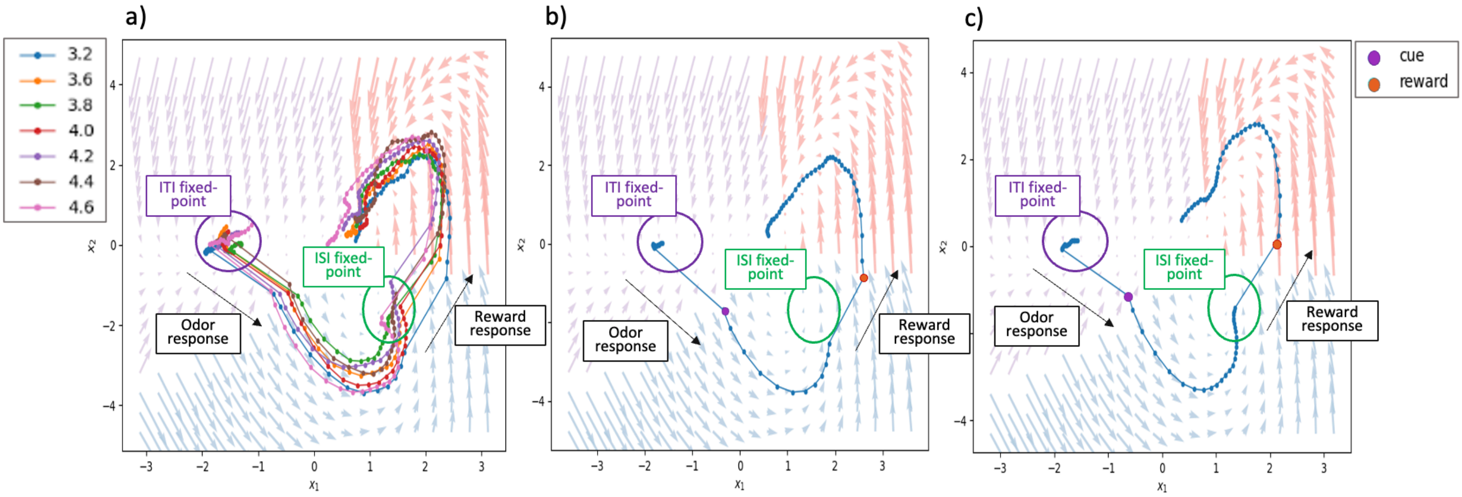


Figure 3.9: Latent trajectories per cue type for best fitted rSLDS model on OFC neural activity. a. Trial-averaged latent trajectory for cue A trials, per reward timing, from 3.2s to 4.8s. Legend corresponds to the time at which the reward happens for each averaged trajectory. **b.** Trial-averaged latent trajectory for cue B trials, for reward at 3.2s. **c.** Trial-averaged latent trajectory for cue C trials, for reward at 4.8s. Flow field corresponds to the most likely dynamics in the at each point, with each color representing one of the 4 linear systems fitted (one per discrete state) and arrows show the direction and rate of change of activity at each point. Fixed-points are identified as the regions of the flow field showing zero-length arrows. For all plots, part of the trajectory before the odor response is the late ITI, part between odor and reward responses is the ISI and part following reward response is the early ITI. ISI fixed-point is identified in the blue linear system, and ITI fixed-point is identified in the purple linear system.

Chapter 4

Discussion

The OFC is thought to play an important role in decision-making in uncertainty. More specifically, it seems to encode for various variables, crucial to belief state computation. Consequently, it has been highlighted as an interesting candidate to hold belief state representation. In this project, we examined neuronal encoding and underlying dynamics in the OFC to investigate the neural basis of belief state computation during a reward learning task with state uncertainty. The variable reward delay task has the advantage of presenting a rather simple paradigm to observe how varying cue-reward intervals change the underlying population dynamics as well as how the transition from one state to the other is computed in the brain.

Delay-tuned neurons and sequential encoding in the OFC Our findings on individual neuron activity strengthen the evidence on the OFC being central to encoding state-inferring variables. OFC was found to encode consistently for odor, delay after odor when expecting a reward delivery and reward. All of these are key elements to the belief state representation in the variable reward delay task. If such a process takes place in the brain, the evidence shown here supports a population-level encoding, where single-neuron information is integrated to infer the belief state.

The sequential activity showed patterns of activation similar to time cells in the hippocampus [90]. Thus, the sub-population of neurons in the sequence could be encoding time after cue. However, the fact that sustained neuronal activity stopped at reward for rewarded trials and the sequential activity was not visible for unrewarded trials indicates that the OFC is not just encoding time. A part of the system, most probably the neurons at the origin of the sequential activity, had to infer the trial value (rewarded or not) as well as the reward delay associated to the observed cue. In line with those observations, latent dynamics of the neural population, obtained from the rSLDS analysis, showed different trajectory velocity in the ISI across odors with different expected delays before reward delivery. Thus, sequential activity could constitute the input to an integrator model predicting value and timing of the reward. In addition, two sub-populations of neurons: the later delay-tuned neurons, with sustained activity from the end of the sequential activity to reward delivery, and the background-tuned neurons, inhibited from cue to reward, could, together be encoding the probability of being in the ISI and the ITI, respectively. The transition of the neural activity from one sub-population to the other would mark the transition between the two macro-states.

Hence, we propose that the OFC neurons individually encode parts of the abstract belief state representation. Neurons could be tuned to different sub-states of the ISI state to be able to track the transitions between the macro-states. Information integrated at the population-level forms the dynamics at the population level. In task 1, transitions are unambiguously signaled by either cue onset or reward onset, making the belief state representation similar to a CSC. Further work should be performed to analyze neural activity in the OFC in task 2. For our proposal to hold, the sequential activity should progressively disperse as probability of state transition increases over the course of the ISI. We expect background-tuned neurons to progressively start firing again while delay-tuned neurons progressively get back to baseline activity before the end of the ISI.

Parallel to the POMDP framework Building on our proposal, we sought to relate neural activity to the POMDP framework (Fig. 1.9). Starkweather and colleagues (2017) provided evidence supporting the TD learning model using belief states operating in the brain [14]. However, for simplicity, it was assumed that the only source of uncertainty in the variable reward delay task is whether the reward was delivered. Consequently, all belief sub-states were weighted similarly across the ISI in task 1 (Fig. 1.7). To account for time uncertainty increasing as time passes, a representation model combining microstimuli and belief state representations was proposed. Belief state representation is sensitive to the task structure. Hence, by conceiving microstimuli as derived from the belief state, the microstimuli shape could vary to adapt to the task structure and encode progression in the ISI. Thus, by combining both the representation would be more flexible than the belief state representation alone. Plus, microstimuli could be thought as the neural realizations of the abstract-state representation implied by the belief-state model and the brain would encode abstract belief state through neurons whose temporal-receptive fields resemble microstimuli [14].

Various results from our analyses reinforce this theory. Neurons part of the sequential activity, both peaking at a specific moment in time or showing sustained activity, actually mirrored the evolution of sub-states in the microstimulus model (Fig. 1.3). As time elapses, the model incorporates neural timing noise that accrues for longer intervals. This fits with our micro-states decoder showing decaying performances as time passed in the ISI. Further increasing the resemblance, latent dynamics of the neural activity showed a variable velocity of the activity trajectory at odor response, depending on the reward delay. It goes in pair with the expectation that the microstimuli adapt their shapes to fit with the task structure.

Subsequent analysis would consist of designing the ISI and reward variable bases in our encoder similarly to microstimuli representation. If the OFC is encoding sub-states as microstimuli, such an encoder should capture more of the neural activity than our current implementation, resembling CSC representation with a ballistic sequence of equally dispersed sub-states. Analysis on task 2 should also be performed.

Neural population activity decoding macro- and micro-states Neural population activity was shown to be sufficient to decode states. The decoder on micro-states performed better on earlier ISI states with accuracy decreasing as time elapsed in the ISI. We can safely reject the possibility that the difference of performances between the early and late states of the ISI were caused by a lack of samples for the later states, due to how we re-balanced the data. Indeed, the performances also degraded in the later states for cue C, in which the number of samples was similar for all states. We rather hypothesize that the difference is due to the sustained activation of neurons in the late ISI. It makes the overall activity more similar from one adjacent state to the next.

In light of the dynamics of the neural population, the micro-states in the early ISI might be more distinct compared to those in the later ISI because their neural activity is not yet stabilized at the fixed-point. Those states consist of the activity transitioning from the ITI to the ISI fixed-point. In that sense, the OFC neural activity contains the dynamics needed to differentiate macro-states: activity is differentiable in both macro-states, with baseline activity for the ITI and neurons sustaining a high activity in the late ISI. Transitions are equivalent to the trajectory going from one macro-state to the other in the dynamical space, and the neural activity at that time is distinct from one sub-state to the next.

An interesting follow-up would be to compare classifiers, trained on only odor-A and odor-C trials respectively, on this late ISI period, in task 2. We expect the prediction to be more accurate after cue C for which the animal knows until expected time of reward that it cannot be in the ITI already. After cue A, the animal is in constant uncertainty about the state until the last possible reward delivery timing happens.

Functional heterogeneity in the OFC New data were recorded on 4 extra animals trained on task 1 and task 2. We did not have the time to analyze them carefully in the time frame of this project. However, until now, neural data was recorded in the lateral part of the OFC. Some of those new sessions recorded neurons in the OFC in different sub-regions, i.e. from the ventro-lateral and medial parts. Preliminary results indicated that neurons recorded in different sub-regions might not be tuned to the same task features as what we observed in this project (Supp. Fig. A.18). Especially neural activity recorded in the ventro-lateral and medial OFC does not seem to show tuning to reward delivery.

And indeed, the medial and lateral OFC have distinct connectivity and such difference has been assumed to have functional implications [18]. Based on a lesion study, Noonan and colleagues (2010) proposed a distinct division between mOFC and lOFC, in term of their function. The mOFC would focus on reward-guided decision making and value comparison. In comparison, the lOFC's role would be in reward-guided learning, encoding the credit assignment problem [53] - i.e. attributing a reward to the corresponding action for learning the value of said-action [91]. However, if those functions were artificially differentiated in order to be investigated, they are not necessarily opposed in natural settings. They might rely on the same states representations: the agent *learns* a task, using credit assignment to improve states representation and can later *take decisions* based on its improved state representation. With this hypothesis, states representation would actually be the link between the two initially distinct functions.

Investigation on the new data needs to be performed to understand the discrepancy in neuronal tuning better. In light of those studies and considering that the task we used is a reward learning task, the task would require mostly lateral OFC rather than medial OFC. Thus, it would make sense that we observe different patterns of activations between those two sub-regions.

Delay-tuned neurons in the M2 Interestingly, we found that the M2 neurons presented delay-tuned neurons, showing sequential activity along the ISI duration, similar to the ones in the OFC. Moreover, our models trained on M2, both the linear regression model (Supp. Fig. A.5) and the linear classification (Supp. Fig. A.9), had performances comparable to the ones trained on the OFC neural activity. However, contrary to neurons in the OFC, no odor-tuned neurons were observed.

This corroborates a previous study on the involvement of M2 in decision-making, and more precisely in the initiation of action [92]. Using a task in which rats had to decide when to abort waiting for a delayed tone, the authors observed two types of neural populations in the M2. Neurons in the first population would ramp to a constant threshold, at rates proportional to the waiting time, while the second population fired in sequences in time until waiting is over. In light of the integration-to-bound mechanism, they associate the first population to the integrator output, identifying the timing of the initial intention to act, while the second would represent the inputs to the integrator.

Similarly, we could relate the neurons temporal alignment observed in the ISI to the input to an integrator model from the cue to the reward. This way, no integration is needed in unrewarded trials, as the animal has no interest in keeping track of time knowing that no reward will follow.

Odor specificity of the belief state representation Both the POMDP presented by Starkweather and colleagues [14] and the following project did not investigate differences in neural encoding across different odors. It could provide insights on the generalization of the encoding taking place in the brain. Multiple possibilities exist. Beliefs for each odor could share the same ITI state representation but otherwise be separated, meaning a same time point after cue in trial would be represented differently depending on the odor. Alternatively, state representation could be as common as possible, relying on the downstream value readout process to differentiate them. Finally, considering that (1) regardless of the odor, the neural network relies on the same dynamics but (2) different odors have different values, we could envision a mixture of the above where a dynamical state representation is common to all odors but the representation differs regarding the value each state encodes.

For simplification, most of our analysis were presented while separating the trial types. Yet, when comparing our findings between trial types, the OFC neurons showed similar activity patterns across odors. For instance the low-dimensional spaces obtained with the PCA analysis for each trial type showed similar geometrical representations from one another (Fig. 3.7). Preliminary investigation also constructed a linear regression model trained on a regression matrix with a unique variable for all the different odors. Performances were found to be as good as the model for which the regression matrix has one variable per cue type (Supp. Fig. A.19, Top). Additionally, a model trained only on odor A trials could predict the spiking rate in the other trial types almost as good as the full linear regression model, and similarly for a model trained on odor C trials only (Supp. Fig. A.19, Bottom). However, the dynamics of the neural activity showed variations in velocity in the ISI, depending on the odor onset. It hints that there is still a difference regarding the value each state encodes.

Hence, preliminary results on odor specificity indicates that a common dynamical state representation that differs regarding the value each state encodes might be the process in place. More analyses should be performed to reinforce this idea.

comments and possible improvements on the rSLDS modeling Using the rSLDS on neural data is a new analytic method and there are still a few points that could be improved for further work and that would have been investigated if time would have allowed it.

First of all, we only displayed the latent space for the best fitted model, out of around a hundred model fittings. The rSLDS optimizatipn process being nonconvex, reinitialization is needed to ensure that the model did not converge to a local mimum rather than a global one. We still investigated the dynamics of the sub-optimal fitted models and they actually showed a great variability. Of interest, some did not succeed in capturing both fixed-points (Supp. Fig. A.15; no clear ISI fixed-point when looking at the flow field). Many also displayed a line attractor dynamic rather than two fixed-points attractors (Supp. Fig. A.17). In that last case, the late ITI (before odor response) and late ISI (before reward response) were found both on the line, each on one side. Hence, even if results might not have been as striking for the other fittings, we want to emphasize that the main findings were always captured.

Second, we could have improved the trajectory of the early ITI not systematically displayed as coming back to the ITI fixed-point. We observed that the trajectory in late ITI even converged back to the ISI fixed-point in some sub-optimal fitted models (Supp. Fig. A.16). This is probably due to the short duration of the early ITI. We saw in our individual neuron analysis that some reward-tuned neurons might take a few seconds to come back to baseline activity after reward. Here, we selected trials, cutting the ITI 4s after reward only, as it is the minimum ITI duration in the trial and we didn't want to have trials overlapping. If the model did not get enough samples to infer the ”end of the loop”, once the reward response is over, it makes sense that it is wrongly estimating the trajectory. Designing the experiment with a longer minimal ITI duration or processing trials with varying lengths could be a solution.

Finally, we investigated more specifically the presence of fixed-points in the two-dimensional plane. In order to implement the model faster, we made the assumption that two latent dimensions were sufficient to describe the system. However, Nair and colleagues performed a factor analysis on the neural data to tune the number of latent dimensions. They took the minimum number of dimensions that explains at least 90% of the observed variance [17]. We also performed such an analysis (Supp. Fig. A.20) and 6 was found to be the minimum number of latent space. As we still find a two-dimensional space in line with what we were expecting, we can consider that our models provided a reasonable approximation of neural activity already. However, further work on the dynamics matrix for instance should be performed on models with a well-tuned number of latent dimensions.

References

1. Niv, Y. Reinforcement learning in the brain. *J. Math. Psychol.* **53,** 139–154. ISSN: 0022-2496 (June 2009).
2. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275,** 1593–1599. ISSN: 0036-8075 (Mar. 1997).
3. Glimcher Paul, W. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 15647–15654. ISSN: 0027-8424 (Sept. 2011).
4. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural Circuitry of Reward Prediction Error. *Annu. Rev. Neurosci.* **40,** 373–394. ISSN: 1545-4126. eprint: 28441114 (July 2017).
5. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (The MIT Press, 1998).
6. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101,** 99–134. ISSN: 0004-3702 (May 1998).
7. Kakade, S. & Dayan, P. Dopamine: Generalization and Bonuses. *Neural networks : the official journal of the International Neural Network Society* **15,** 549–59. ISSN: 0893-6080 (June 2002).
8. Daw, N. D., Courville, A. C. & Touretzky, D. S. Representation and timing in theories of the dopamine system. *Neural Comput.* **18,** 1637–1677. ISSN: 0899-7667. eprint: 16764517 (July 2006).
9. Rao, R. P. N. Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. *Front. Comput. Neurosci.* **0.** ISSN: 1662-5188 (2010).
10. Takahashi, Y. K., Langdon, A. J., Niv, Y. & Schoenbaum, G. Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. *Neuron* **91,** 182–193. ISSN: 0896-6273 (July 2016).
11. Lak, A., Nomoto, K., Keramati, M., Sakagami, M. & Kepecs, A. Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Curr. Biol.* **27,** 821–832. ISSN: 0960-9822 (Mar. 2017).
12. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System. *Neural Comput.* **20,** 3034–3054. ISSN: 0899-7667 (Dec. 2008).
13. Gershman, S. J., Moustafa, A. A. & Ludvig, E. A. Time representation in reinforcement learning models of the basal ganglia. *Front. Comput. Neurosci.* **0.** ISSN: 1662-5188 (2014).
14. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20,** 581–589. ISSN: 1546-1726 (Apr. 2017).
15. Babayan, B. M., Uchida, N. & Gershman, S. J. Belief state representation in the dopamine system. *Nat. Commun.* **9,** 1–10. ISSN: 2041-1723 (May 2018).
16. Starkweather, C. K., Gershman, S. J. & Uchida, N. The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. *Neuron* **98,** 616–6296. ISSN: 1097-4199. eprint: 29656872 (May 2018).
17. Nair, A. *et al.* An approximate line attractor in the hypothalamus that encodes an aggressive internal state. *bioRxiv,* 2022.04.19.488776. eprint: 2022.04.19.488776. https://doi.org/10.1101/2022.04.19.488776 (Apr. 2022).
18. Schuck, N. W., Wilson, R. & Niv, Y. in *Goal-Directed Decision Making* 259–278 (Academic Press, Cambridge, MA, USA, Jan. 2018). ISBN: 978-0-12-812098-9.
19. Bellman, R. A Markovian Decision Process on JSTOR. *Journal of Mathematics and Mechanics* **6,** 679–684. https://www.jstor.org/stable/24900506 (1957).
20. Sutton, R. S. & Barto, A. G. *Time-derivative models of Pavlovian reinforcement.* In M. Gabriel J. Moore (Eds.), 497–537 (The MIT Press, Cambridge, MA, 1990).
21. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3,** 9–44. ISSN: 1573-0565 (Aug. 1988).

22. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16,** 1936–1947. ISSN: 0270-6474 (Mar. 1996).

23. Hollerman, J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1,** 304–309. ISSN: 1546-1726 (Aug. 1998).

24. Rescorla, R. A. "Configural" conditioning in discrete-trial bar pressing. *Journal of Comparative and Physiological Psychology* **79,** 307–317 (1972).

25. Bayer, H. M. & Glimcher, P. W. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* **47,** 129–141. ISSN: 0896-6273 (July 2005).

26. Mirenowicz, J. & Schultz, W. Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.* **72,** 1024–1027. ISSN: 0022-3077. eprint: 7983508 (Aug. 1994).

27. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482,** 85–88. ISSN: 1476-4687 (Feb. 2012).

28. Eshel, N. *et al.* Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* **525,** 243–246. ISSN: 1476-4687. eprint: 26322583 (Sept. 2015).

29. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19,** 479–486. ISSN: 1546-1726 (Mar. 2016).

30. Roesch, M. R., Calu, D. J. & Schoenbaum, G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.* **10,** 1615–1624. ISSN: 1097-6256. eprint: 18026098 (Dec. 2007).

31. Pan, W.-X., Schmidt, R., Wickens, J. R. & Hyland, B. I. Dopamine Cells Respond to Predicted Events during Classical Conditioning: Evidence for Eligibility Traces in the Reward-Learning Network. *J. Neurosci.* **25,** 6235–6242. ISSN: 0270-6474 (June 2005).

32. D'Ardenne, K., McClure, S. M., Nystrom, L. E. & Cohen, J. D. BOLD Responses Reflecting Dopaminergic Signals in the Human Ventral Tegmental Area. *Science* **319,** 1264–1267. ISSN: 0036-8075 (Feb. 2008).

33. Hart, A. S., Rutledge, R. B., Glimcher, P. W. & Phillips, P. E. M. Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term. *J. Neurosci.* **34,** 698–704. ISSN: 0270-6474 (Jan. 2014).

34. Stuber, G. D. *et al.* Reward-Predictive Cues Enhance Excitatory Synaptic Strength onto Midbrain Dopamine Neurons. *Science* **321,** 1690–1692. ISSN: 0036-8075 (Sept. 2008).

35. Balsam, P. D. & Gallistel, C. R. Temporal maps and informativeness in associative learning. *Trends Neurosci.* **32,** 73–78. ISSN: 0166-2236 (Feb. 2009).

36. Starkweather, C. K. & Uchida, N. Dopamine reward prediction errors: The interplay between experiments and theory. *The Cognitive Neuroscience* (2020).

37. Rao, R. P. N. *Neural Models of Bayesian Belief Propagation* ISBN: 978-0-26204238-3 (Jan. 2007).

38. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA, 2009).

39. Pearl, J. *Probabilistic Reasoning in Intelligent Systems* ISBN: 978-0-08-051489-5 (Morgan Kaufmann, 1988).

40. Stalnaker, T. A., Berg, B., Aujla, N. & Schoenbaum, G. Cholinergic Interneurons Use Orbitofrontal Input to Track Beliefs about Current State. *J. Neurosci.* **36,** 6242–6257. ISSN: 0270-6474 (June 2016).

41. Vertechi, P. *et al.* Inference-Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas. *Neuron* **106,** 166–176.e6. ISSN: 0896-6273 (Apr. 2020).

42. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron* **81,** 267–279. ISSN: 0896-6273 (Jan. 2014).

43. Vardy, E. *et al.* A New DREADD Facilitates the Multiplexed Chemogenetic Interrogation of Behavior. *Neuron* **86,** 936–946. ISSN: 0896-6273 (May 2015).

44. Cavada, C., Compay, T., Tejedor, J., Cruz-Rizzolo, R. J. & Reinoso-Surez, F. The Anatomical Connections of the Macaque Monkey Orbitofrontal Cortex. A Review. *Cereb. Cortex* **10,** 220–242. ISSN: 1047-3211 (Mar. 2000).

45. Kahnt, T., Chang, L. J., Park, S. Q., Heinzle, J. & Haynes, J.-D. Connectivity-Based Parcellation of the Human Orbitofrontal Cortex. *J. Neurosci.* **32,** 6240–6250. ISSN: 0270-6474 (May 2012).

46. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455,** 227–231. ISSN: 1476-4687 (Sept. 2008).

47. Roesch, M. R., Taylor, A. R. & Schoenbaum, G. Encoding of Time-Discounted Rewards in Orbitofrontal Cortex Is Independent of Value Representation. *Neuron* **51,** 509–520. ISSN: 0896-6273 (Aug. 2006).

48. Padoa-Schioppa, C. & Assad, J. A. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat. Neurosci.* **11,** 95–102. ISSN: 1097-6256. eprint: 18066060 (Jan. 2008).

49. Schoenbaum, G. & Eichenbaum, H. Information coding in the rodent prefrontal cortex. I. Single-neuron activity in orbitofrontal cortex compared with that in pyriform cortex. *J. Neurophysiol.* (Aug. 1995).

50. Schoenbaum, G., Nugent, S. L., Saddoris, M. P. & Setlow, B. Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. *Neuroreport* **13,** 885–890. ISSN: 0959-4965. https://journals.lww.com/neuroreport/Abstract/2002/05070/Orbitofrontal_lesions_in_rats_impair_reversal_but.30.aspx (May 2002).

51. Schoenbaum, G., Setlow, B., Saddoris, M. P. & Gallagher, M. Encoding Predicted Outcome and Acquired Value in Orbitofrontal Cortex during Cue Sampling Depends upon Input from Basolateral Amygdala. *Neuron* **39,** 855–867. ISSN: 0896-6273 (Aug. 2003).

52. Sharpe, M. J. & Schoenbaum, G. Back to basics: Making predictions in the orbitofrontal–amygdala circuit. *Neurobiol. Learn. Mem.* **131,** 201–206. ISSN: 1074-7427 (May 2016).

53. Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H. & Rushworth, M. F. S. Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. *Neuron* **65,** 927 (Mar. 2010).

54. Tsujimoto, S., Genovesio, A. & Wise, S. P. Frontal pole cortex: encoding ends at the end of the endbrain. *Trends in Cognitive Sciences* **15,** 169–176. ISSN: 1364-6613 (Apr. 2011).

55. Brown, T. I., Ross, R. S., Keller, J. B., Hasselmo, M. E. & Stern, C. E. Which Way Was I Going? Contextual Retrieval Supports the Disambiguation of Well Learned Overlapping Navigational Routes. *J. Neurosci.* **30,** 7414–7422. ISSN: 0270-6474 (May 2010).

56. Nee, D. E. & Brown, J. W. Rostral-Caudal Gradients of Abstraction Revealed by Multi-Variate Pattern Analysis of Working Memory. *Neuroimage* **63,** 1285 (Nov. 2012).

57. Ramus, S. J. & Eichenbaum, H. Neural Correlates of Olfactory Recognition Memory in the Rat Orbitofrontal Cortex. *J. Neurosci.* **20,** 8199–8208. ISSN: 0270-6474 (Nov. 2000).

58. Bradfield, L. A., Dezfouli, A., van Holstein, M., Chieng, B. & Balleine, B. W. Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. *Neuron* **88,** 1268–1280. ISSN: 1097-4199. eprint: 26627312 (Dec. 2015).

59. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160,** 106–154. ISSN: 0022-3751 (Jan. 1962).

60. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551,** 232–236. ISSN: 1476-4687 (Nov. 2017).

61. Arandia-Romero, I., Nogueira, R., Mochol, G. & Moreno-Bote, R. What can neuronal populations tell us about cognition? *Curr. Opin. Neurobiol.* **46,** 48–57. ISSN: 0959-4388 (Oct. 2017).

62. Holdgraf, C. R. *et al.* Encoding and Decoding Models in Cognitive Electrophysiology. *Front. Syst. Neurosci.* **0.** ISSN: 1662-5137 (2017).

63. Kiani, R., Cueva, C. J., Reppas, J. B. & Newsome, W. T. Dynamics of Neural Population Responses in Prefrontal Cortex Indicate Changes of Mind on Single Trials. *Curr. Biol.* **24,** 1542–1547. ISSN: 0960-9822 (July 2014).

64. Rich, E. L. & Wallis, J. D. Decoding subjective decisions from orbitofrontal cortex. *Nat. Neurosci.* **19,** 973–980. ISSN: 1546-1726 (July 2016).

65. Nogueira, R. *et al.* Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. *Nat. Commun.* **8,** 1–13. ISSN: 2041-1723 (Mar. 2017).

66. Ziv, Y. *et al.* Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16,** 264–266. ISSN: 1546-1726 (Mar. 2013).

67. Druckmann, S. & Chklovskii, D. B. Neuronal Circuits Underlying Persistent Representations Despite Time Varying Activity. *Curr. Biol.* **22,** 2095–2103. ISSN: 0960-9822 (Nov. 2012).

68. Doya, K. Universality of Fully-Connected Recurrent Neural Networks. https://www.semanticscholar.org/paper/Universality-of-Fully-Connected-Recurrent-Neural-Doya/0724c8219db73af52ecd45cc6%20afeba3c12e7fe57#paper-header (1993).

69. Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S. & Sussillo, D. Universality and individuality in neural dynamics across large populations of recurrent networks. *arXiv.* eprint: 1907.08549 (July 2019).

70. Cho, K. *et al.* Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv.* eprint: 1406.1078 (June 2014).

71. Paszke, A. *et al.* in *Advances in Neural Information Processing Systems 32* (eds Wallach, H. *et al.*) 8024–8035 (Curran Associates, Inc., 2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

72. Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* **7,** 1–10. ISSN: 2041-1723 (Nov. 2016).

73. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J. & Kurth-Nelson, Z. Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron* **107,** 603–616. ISSN: 0896-6273 (Aug. 2020).

74. Council, N. R. *Guide for the Care and Use of Laboratory Animals 8th edn* (The National Academies Press, 2011).

75. MouseLand. *Kilosort* [Online; accessed 22. Aug. 2022]. Aug. 2022. https://github.com/MouseLand/Kilosort.

76. *Manual Clustering Practical Guide (by S. Lenzi and N. Steinmetz) - phy* [Online; accessed 22. Aug. 2022]. Jan. 2021. https://phy.readthedocs.io/en/latest/sorting_user_guide.

77. petersaj. *AP_histology* [Online; accessed 29. Jul. 2022]. July 2022. https://github.com/petersaj/AP_histology.

78. Wang, Q. *et al.* The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181,** 936–953.e20. ISSN: 0092-8674 (May 2020).

79. *Visual Coding – Neuropixels — Allen SDK dev documentation* [Online; accessed 25. Jul. 2022]. https://allensdk.readthedocs.io/en/latest/visual_coding_neuropixels.html.

80. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17,** 261–272 (2020).

81. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585,** 357–362. https://doi.org/10.1038/s41586-020-2649-2 (Sept. 2020).

82. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

83. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)* **135,** 370–384. ISSN: 0035-9238 (May 1972).

84. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320. ISSN: 1369-7412 (Apr. 2005).

85. Jas, M. *et al.* Pyglmnet : Python implementation of elastic-net regularized generalized linear models. *JOURNAL OF OPEN SOURCE SOFTWARE* **5.** ISSN: 2475-9066. https://biblio.ugent.be/publication/8650816 (2020).

86. Linderman, S. *et al.* in *Artificial Intelligence and Statistics* 914–922 (PMLR, Apr. 2017). http://proceedings.mlr.press/v54/linderman17a.html.

87. Linderman, S., Antin, B., Zoltowski, D. & Glaser, J. *SSM: Bayesian Learning and Inference for State Space Models* version 0.0.1. Oct. 2020. https://github.com/lindermanlab/ssm.

88. Linderman, S., Nichols, A., Blei, D., Zimmer, M. & Paninski, L. Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in C. elegans. *bioRxiv,* 621540. eprint: 621540. https://doi.org/10.1101/621540 (Apr. 2019).

89. Gottfried, J. A., Winston, J. S. & Dolan, R. J. Dissociable Codes of Odor Quality and Odorant Structure in Human Piriform Cortex. *Neuron* **49,** 467–479. ISSN: 0896-6273 (Feb. 2006).

90. Eichenbaum, H. Time cells in the hippocampus: a new dimension for mapping memories. *Nat. Rev. Neurosci.* **15,** 732–744. ISSN: 1471-0048 (Nov. 2014).

91. Noonan, M. P. *et al.* Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 20547–20552. ISSN: 0027-8424 (Nov. 2010).

92. Murakami, M., Vicente, M. I., Costa, G. M. & Mainen, Z. F. Neural antecedents of self-initiated actions in secondary motor cortex. *Nat. Neurosci.* **17,** 1574–1582. ISSN: 1546-1726 (Nov. 2014).

Appendix

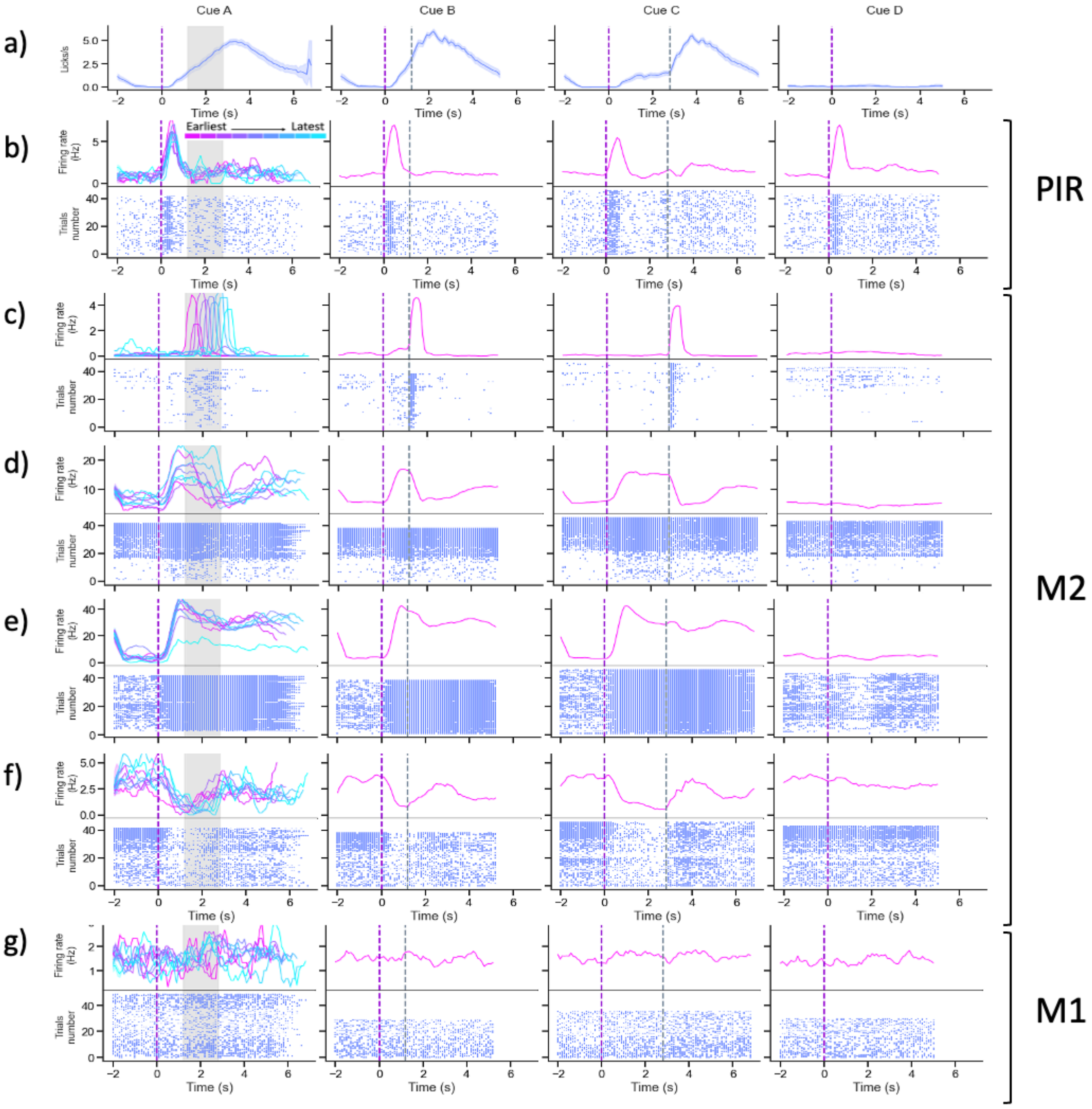


Figure A.1: Licking behavior and example PIR, M2 and M1 neurons. Trial types are separated. Curves correspond to trial-averaged lick count and firing rates, smoothed with a 1s-window 0.1-step moving average. Time zero corresponds to odor onset. The cue onset is indicated with a purple dashed line and reward onset with either a grey area or dashed line. **a.** Lick count per second. Shaded area corresponds to the s.e.m.. **b-g** Example individual neurons in PIR, M2 and M1. Upper plot is the trial-averaged firing rate (Hz) per cue. Lower plot is the spiking raster per trial type, with trials ordered from bottom to top. For cue A, the trial-averaged firing rate are distinctly calculated for each reward timing and displayed from pink to blue with pink being the earliest possible reward (3.2s) and light blue the latest possible reward (4.8s). **b.** Example PIR odor-tuned neuron, activated at odor onset, regardless of the reward timing and even for unrewarded trials (odor D). **c-f** Example M2 neurons. **c.** Example reward-tuned neuron, bursting at specific reward timing in the trial. We see a clear ordering of reward activation in cue A trials from earlier to later reward timings. We do not see activity for odor D. **d.** Example delay-tuned neuron, activated at the same time point regardless of the reward timing, and inhibited at reward. We do not see such a pattern for odor D trials. **e.** Example long delay-tuned neuron, activated at the same time point regardless of the reward timing and staying activated until a fixed time after reward before gradually going back to background activity. No such pattern of activity is visible for odor D trials. **g.** Example M1 neuron, showing no specific tuning to any of the events in the trial.

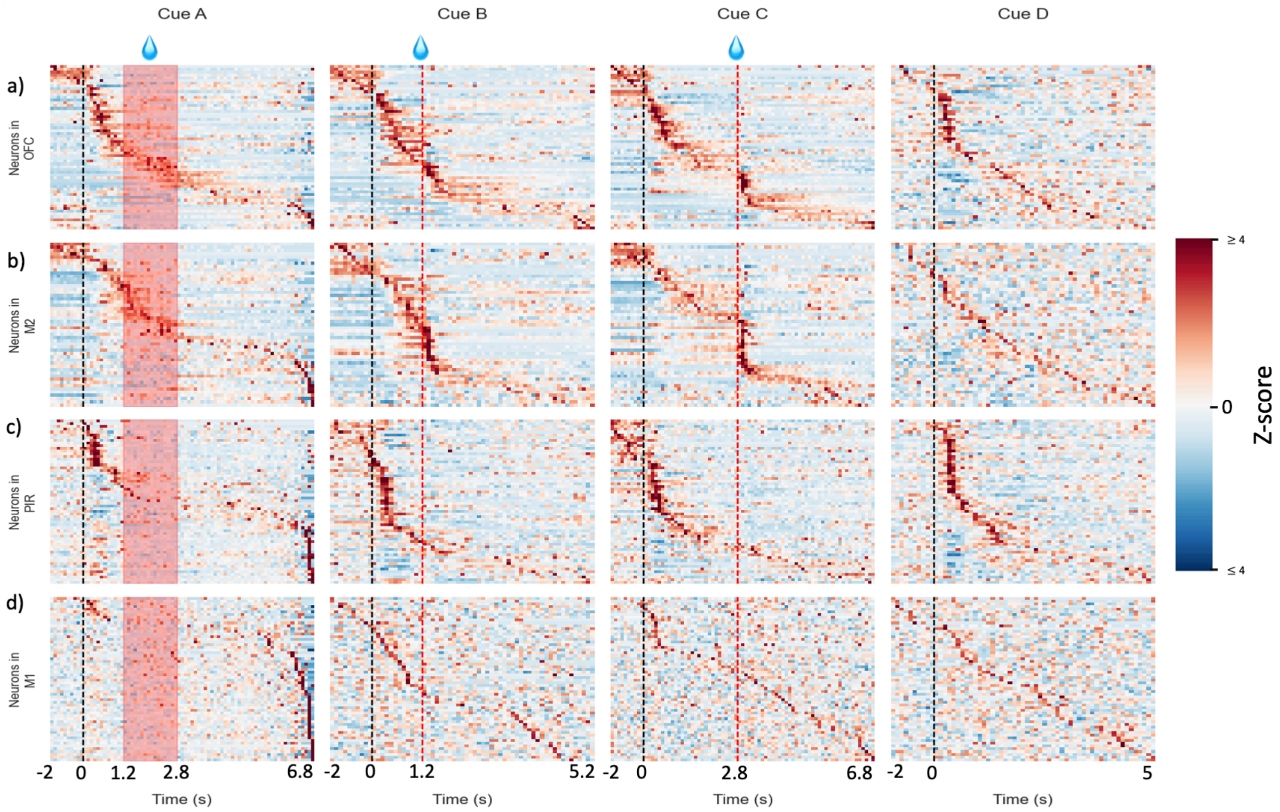


Figure A.2: Training activity maps with neurons ordered per peak activity timing. For each brain region (rows), trial types are separated (columns) and the trial-averaged activity per neuron was computed. Time zero corresponds to odor onset, indicated by a black dashed line. The reward onset is indicated by either a red shaded area or dashed line. For each brain region, neurons are sorted based on the timing of their peak responses from 1s before cue onset to 2s after reward onset. Ordering was computed on half of the trials available for each trial type and we display the - ordered - training set. For each region, we display neuron activity for respectively $n = 59, 51, 60, 60$ neurons for **a.** OFC; **b.** M2; **c.** PIR; and **d.** M1. Color code on the z-score values, in the color bar is bounded from -4 to 4, with values above 4 or below -4 thresholded to those bounds.

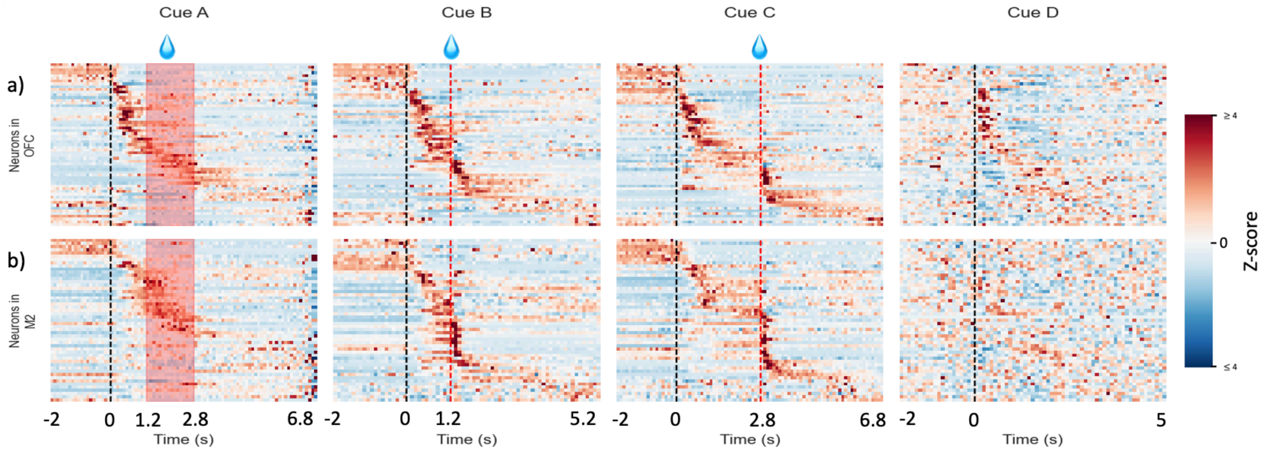


Figure A.3: Activity maps with neurons ordered per peak response timing for a different mouse. For each brain region (rows), trial types are separated (columns) and the trial-averaged activity per neuron was computed. Time zero corresponds to odor onset, indicated by a black dashed line. The reward onset is indicated by either a red shaded area or dashed line. For each brain region, neurons are sorted based on the timing of their peak responses from 1s before cue onset to 2s after reward onset. We display averaged activity on the other half on a larger time window from 2s before cue onset to 4s after reward onset. Neurons are sorted using the ordering found on training activity, from top to bottom. For each region, we display neuron activity for respectively $n = 49, 60$ neurons for **a.** OFC; and **b.** M2. Color code on the z-score values, in the color bar, is bounded from -4 to 4, with values above 4 or below -4 thresholded to those bounds.

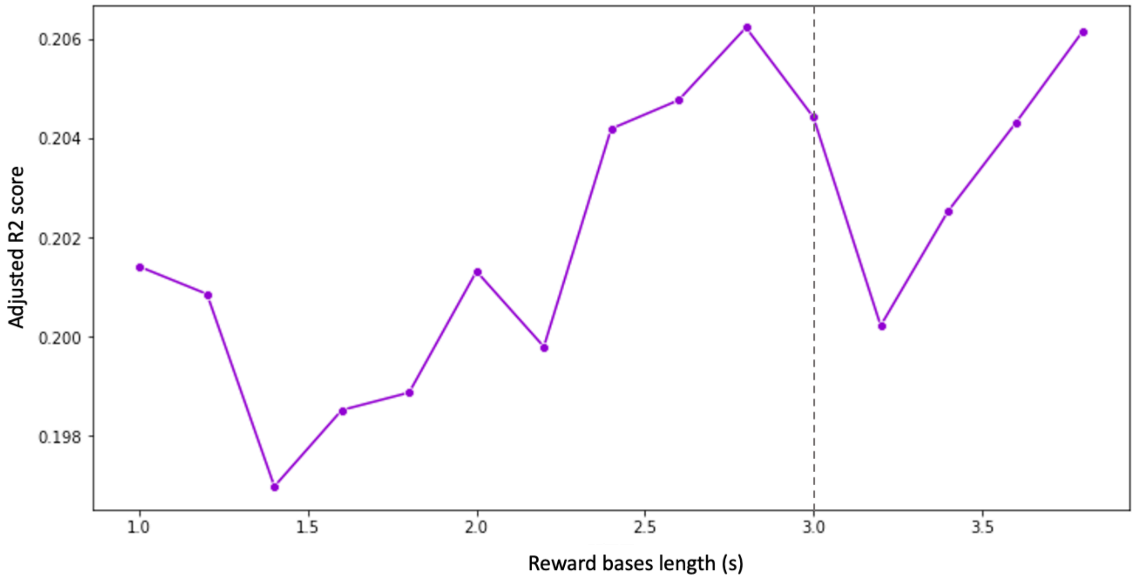


Figure A.4: Grid-search on the length of the reward bases of the linear regression. The linear regression model was fitted with reward bases varying from 1s to 4s (minimum duration of the early ITI, to avoid overlapping with next trial) after onset. We selected 3s because we wanted a round number, close enough to the length of the ISI bases (4.8s) and asserted that performances were good enough compared to other lengths. Best performances were reached for lengths of 2.8s and 4s.

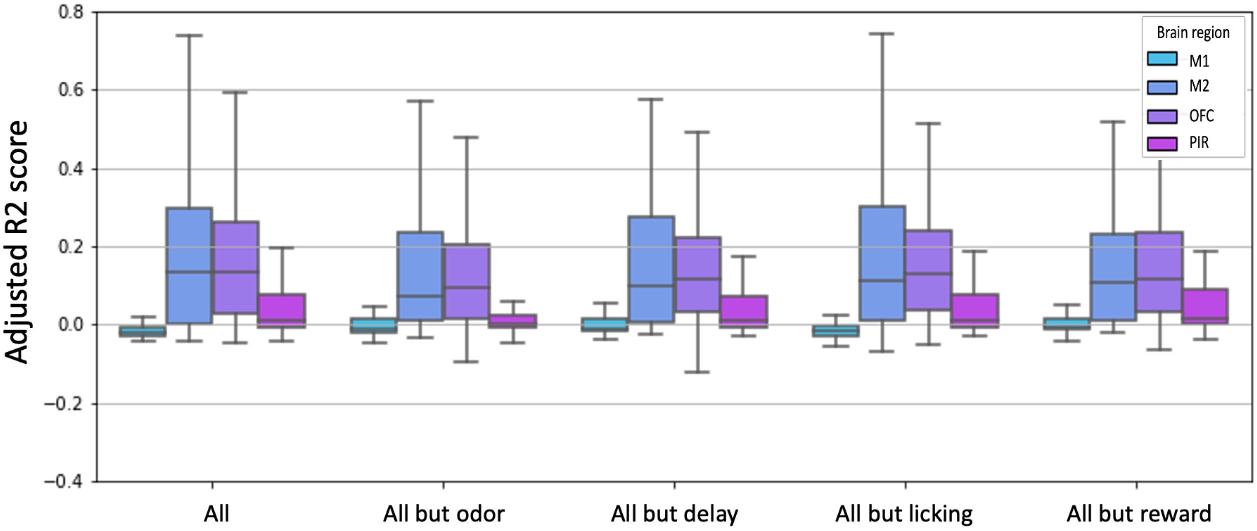


Figure A.5: Distribution of the adjusted R^2-square score across neurons for the LSTSQ linear regression model, trained for full and regressed regression matrices. Left: Model containing all the variables tested, i.e. licking, reward, odor and delay. Then, **Right:** each of the following models are reduced models in which one of the variables was removed. Consequently, the importance of the removed variable is correlated to the performance difference to the full model. Middle line of the box plots is the median score across neurons in the region. We do not display the outliers.

Discrete states	ELBO ($\times 10^6$)
K = 1	-1.057
K = 2	-0.995
K = 3	-1.110
K = 4	-0.67
K = 5	-1.103
K = 6	-1.045

Table A.1: Grid-search optimization of the number of discrete states K in the rSLDS model. For $K \in [1, 6]$, we reinitialized/retrained each model 5 times, compared the average model performances (ELBO) and selected the best K.

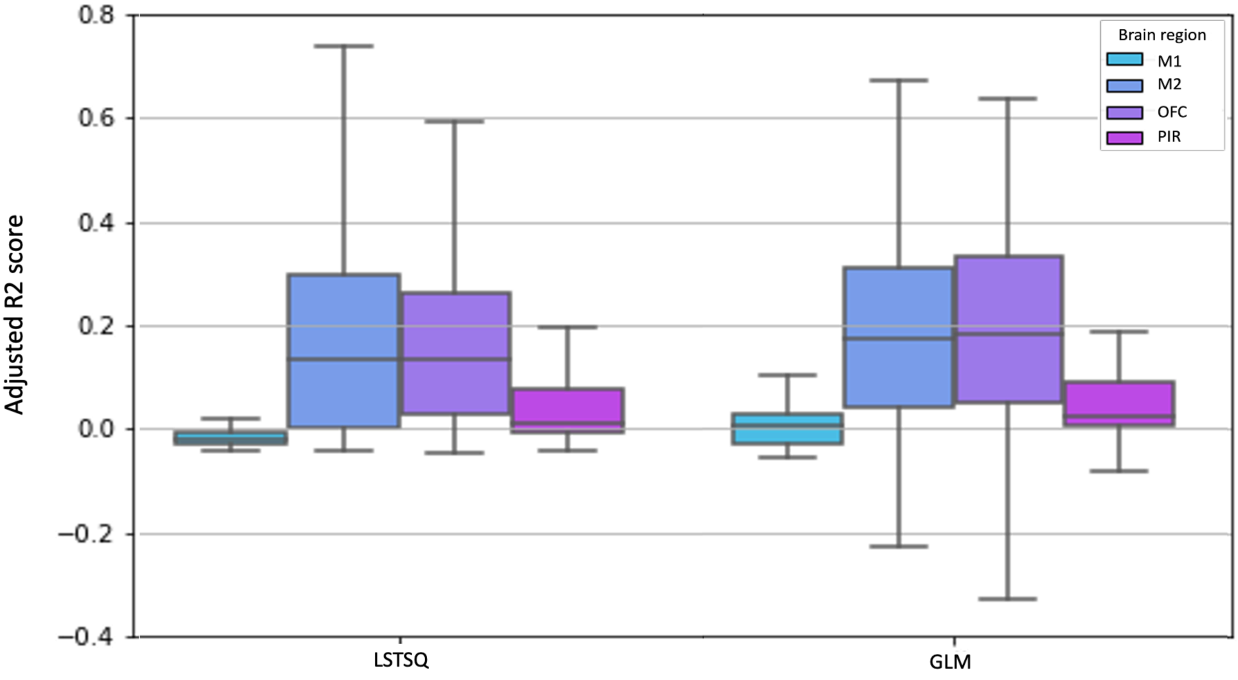


Figure A.6: Performances comparison between GLM model and linear regressor model (LSTSQ). Performances of both models trained on a regression matrix containing all the regressors (ISI, reward, licking) were compared using the adjusted R^2-square score. We see that performances did not get significantly better using the computationally more heavy GLM model. We decided to keep the LSTSQ model as it does not require hyperameters fitting. It took us around 6h to tune one model (one regression matrix format for one region) with the GLM for around 60 neurons, while we could fit all the LSTSQ models (for the five regression matrix combinations tested in this project and all four regions studied) in the same amount of time, while performing equally good.

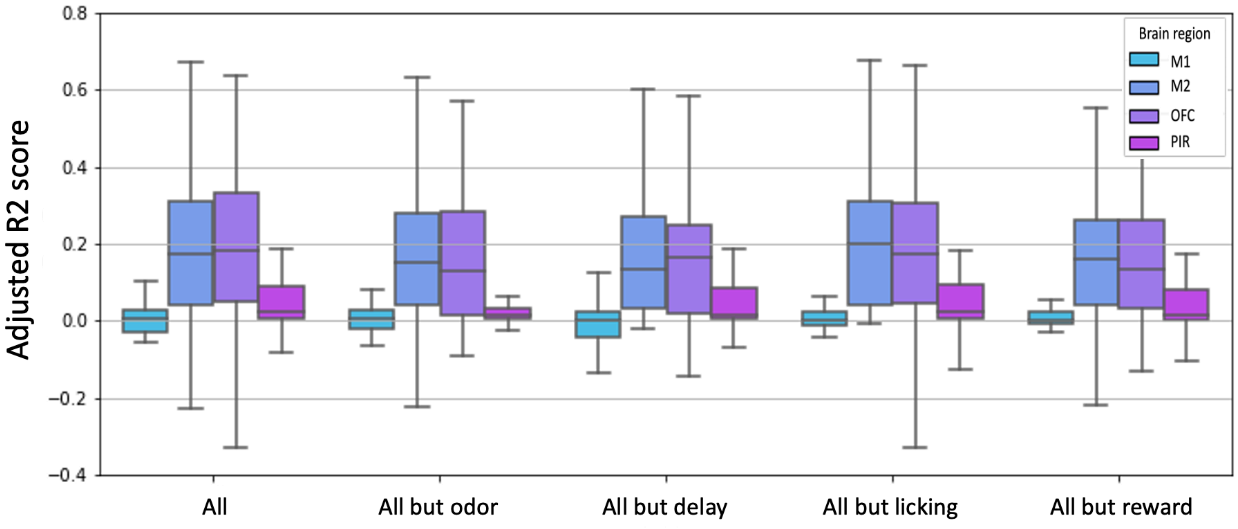


Figure A.7: Distribution of the adjusted R^2-square score across neurons for the GLM model, trained for each regression matrix and each region. Left: Model containing all the variables tested, i.e. licking, reward, odor and delay. Then, **Right:** each of the following models are reduced models in which one of the variables was removed. Consequently, the importance of the removed variable is correlated to the performance difference to the full model. Middle line of the box plots is the median score across neurons in the region. We do not display the outliers.

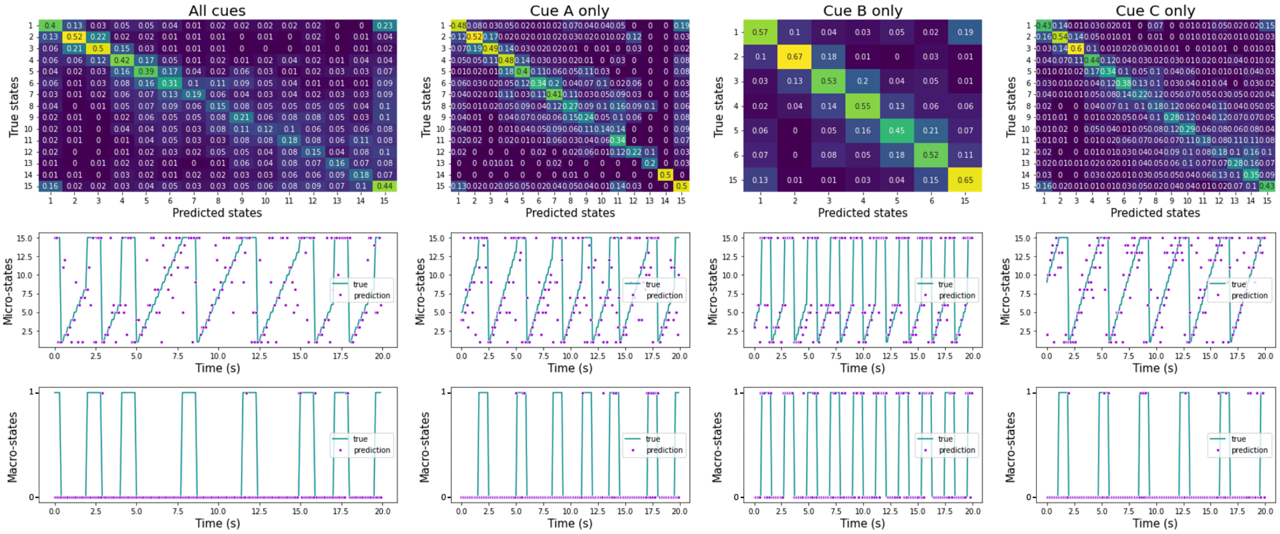


Figure A.8: State classifier performances from neuron population in the PIR cortex. Per column, we trained the decoders on (1) all rewarded cues, (2) only cue A trials, (3) only cue B trials, and (4) only cue C trials. The same analysis was performed for all groups. **(Top)** Confusion matrix normalized by number of occurrences per state (each cell is divided by the sum on values in its row). Values on the diagonal correspond to the sensitivity ($\frac{TP}{TP+FN}$) for each state. Data evaluated is the testing set. **(Middle)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *micro-state* classifier. **(Bottom)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *macro-state* classifier. As one session mostly consists of ITI, the states were first manually re-balanced in the dataset used for training/testing (same number of ITI samples as the maximum available number of samples in an ISI state. Validation data, used for plots in the middle and bottom rows correspond to 6 trials (20s; 200 bins) for 0.2s prior to odor onset to 0.2s after reward onset.

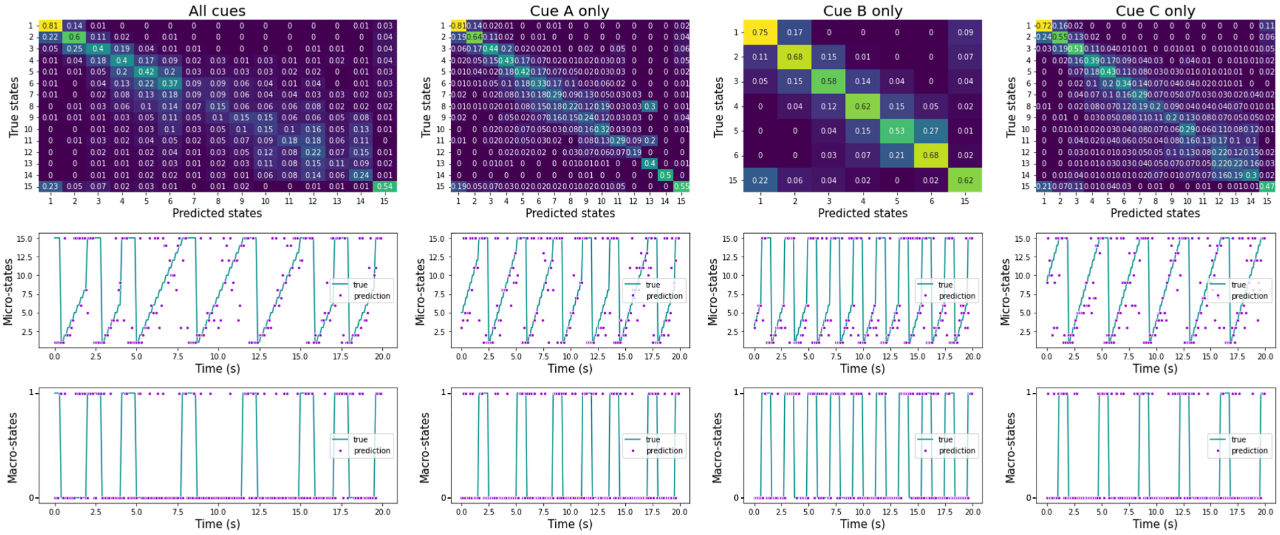


Figure A.9: State decoder performances from neuron population in the M2. Per column, we trained the decoders on (1) all rewarded cues, (2) only cue A trials, (3) only cue B trials, and (4) only cue C trials. The same analysis was performed for all groups. **(Top)** Confusion matrix normalized by number of occurrences per state (each cell is divided by the sum on values in its row). Values on the diagonal correspond to the sensitivity ($\frac{TP}{TP+FN}$) for each state. Data evaluated is the testing set. **(Middle)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *micro-state* classifier. **(Bottom)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *macro-state* classifier. As one session mostly consists of ITI, the states were first manually re-balanced in the dataset used for training/testing (same number of ITI samples as the maximum available number of samples in an ISI state. Validation data, used for plots in the middle and bottom rows correspond to 6 trials (20s; 200 bins) for 0.2s prior to odor onset to 0.2s after reward onset.

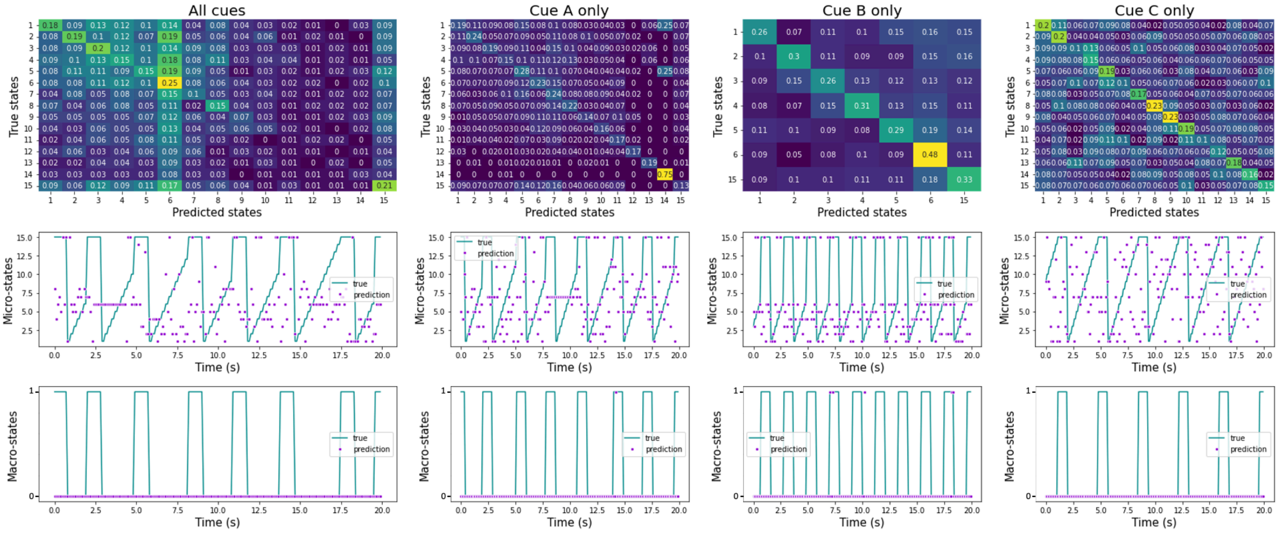


Figure A.10: State decoder performances from neuron population in the M1. Per column, we trained the decoders on (1) all rewarded cues, (2) only cue A trials, (3) only cue B trials, and (4) only cue C trials. The same analysis was performed for all groups. **(Top)** Confusion matrix normalized by number of occurrences per state (each cell is divided by the sum on values in its row). Values on the diagonal correspond to the sensitivity ($\frac{TP}{TP+FN}$) for each state. Data evaluated is the testing set. **(Middle)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *micro-state* classifier. **(Bottom)** Most probable state through time (purple dots) against true state (green) on an example snippet of the session. Most probable state was evaluated by taking the state with the maximum probability for each time point for the *macro-state* classifier. As one session mostly consists of ITI, the states were first manually re-balanced in the dataset used for training/testing (same number of ITI samples as the maximum available number of samples in an ISI state. Validation data, used for plots in the middle and bottom rows correspond to 6 trials (20s; 200 bins) for 0.2s prior to odor onset to 0.2s after reward onset.

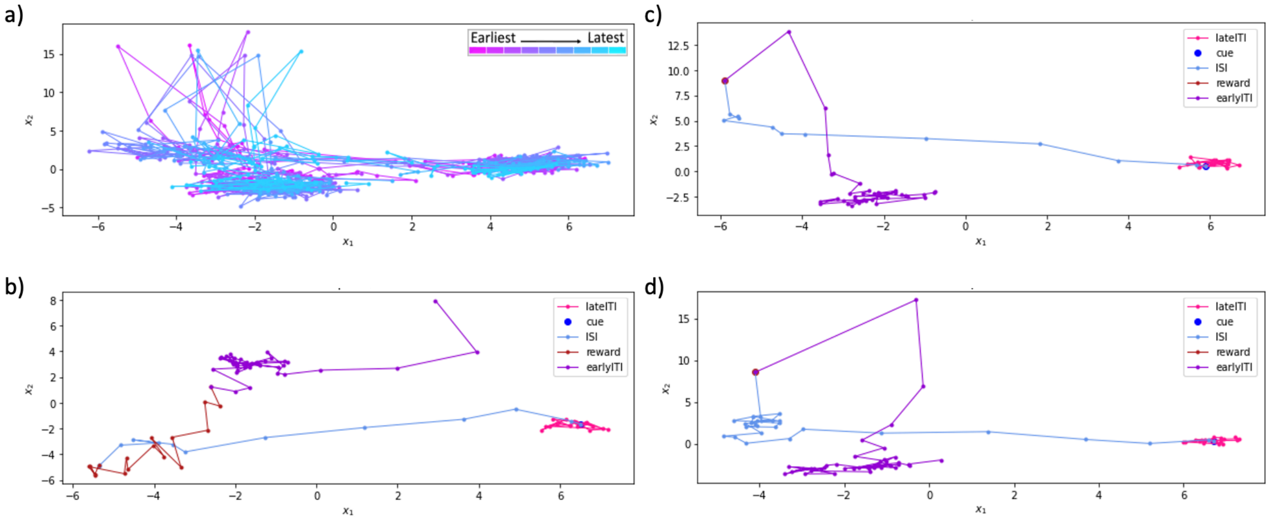


Figure A.11: Dimensionality-reduction of the trial-averaged neural activity per trial type in the M2. PCA was applied on trial-averaged neural activity. We displayed the resulting two first principal components. **b-d** Low-dimensional representation of the trial-average neural activity per trial. Late ITI (pre-cue) in pink, cue in dark blue, ISI in blue, reward in red (either point or line), early ITI in violet. For **a.** odor-A trials, computed distinctively per reward type. Purple trajectories represent earlier rewards trials (from 3.2s) and light blue trajectories are later rewards trials (up to 4.2s). Trajectories are noisier than the other plots as the number of samples to fit the PCA were lower; **b.** odor-A trials, reward timings taken all together (from 3.2s to 4.8s); **c.** odor-B trials (reward at 3.2s); **d.** odor-C trials (reward at 4.8s).

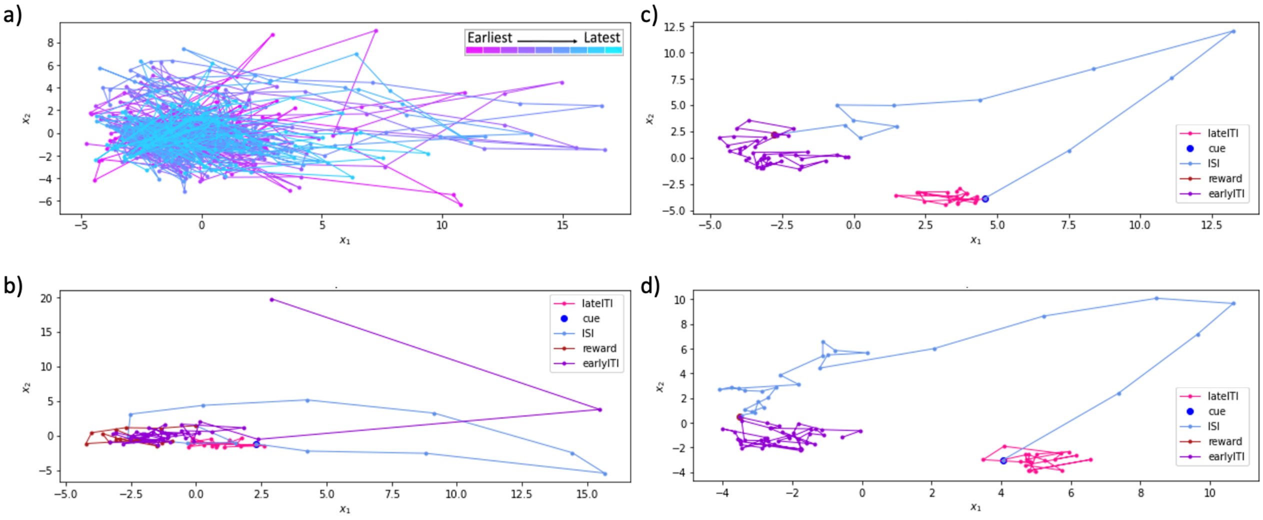


Figure A.12: Dimensionality-reduction of the trial-averaged neural activity per trial type in the PIR. PCA was applied on trial-averaged neural activity. We displayed the resulting two first principal components. **b-d** Low-dimensional representation of the trial-average neural activity per trial. Late ITI (pre-cue) in pink, cue in dark blue, ISI in blue, reward in red (either point or line), early ITI in violet. For **a.** odor-A trials, computed distinctively per reward type. Purple trajectories represent earlier rewards trials (from 3.2s) and light blue trajectories are later rewards trials (up to 4.2s). Trajectories are noisier than the other plots as the number of samples to fit the PCA were lower; **b.** odor-A trials, reward timings taken all together (from 3.2s to 4.8s); **c.** odor-B trials (reward at 3.2s); **d.** odor-C trials (reward at 4.8s).

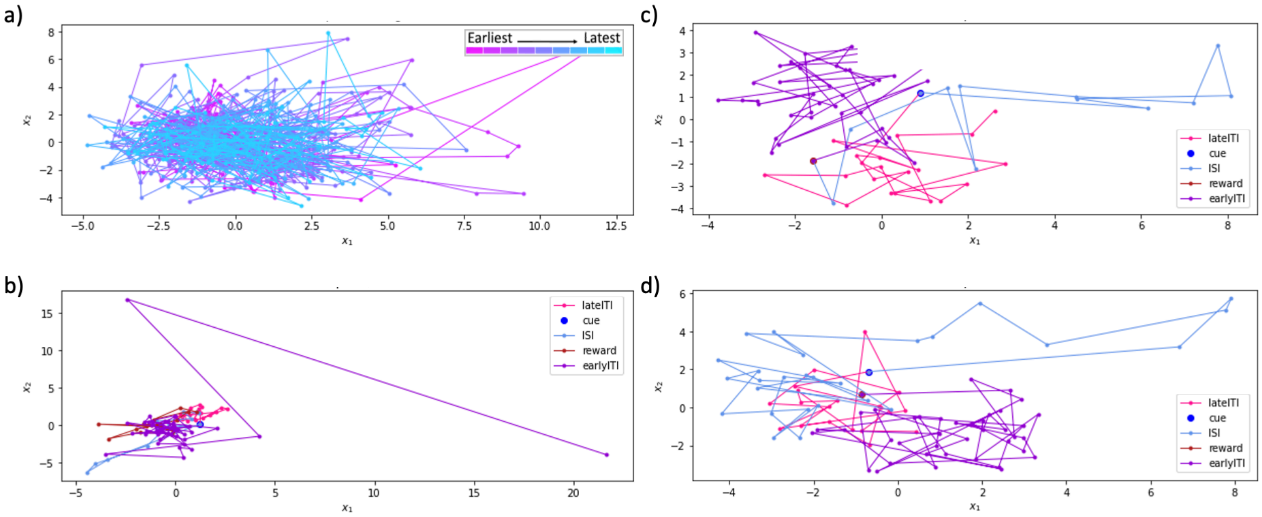


Figure A.13: Dimensionality-reduction of the trial-averaged neural activity per trial type in the M1. PCA was applied on trial-averaged neural activity. We displayed the resulting two first principal components. **b-d** Low-dimensional representation of the trial-average neural activity per trial. Late ITI (pre-cue) in pink, cue in dark blue, ISI in blue, reward in red (either point or line), early ITI in violet. For **a.** odor-A trials, computed distinctively per reward type. Purple trajectories represent earlier rewards trials (from 3.2s) and light blue trajectories are later rewards trials (up to 4.2s). Trajectories are noisier than the other plots as the number of samples to fit the PCA were lower; **b.** odor-A trials, reward timings taken all together (from 3.2s to 4.8s); **c.** odor-B trials (reward at 3.2s); **d.** odor-C trials (reward at 4.8s).

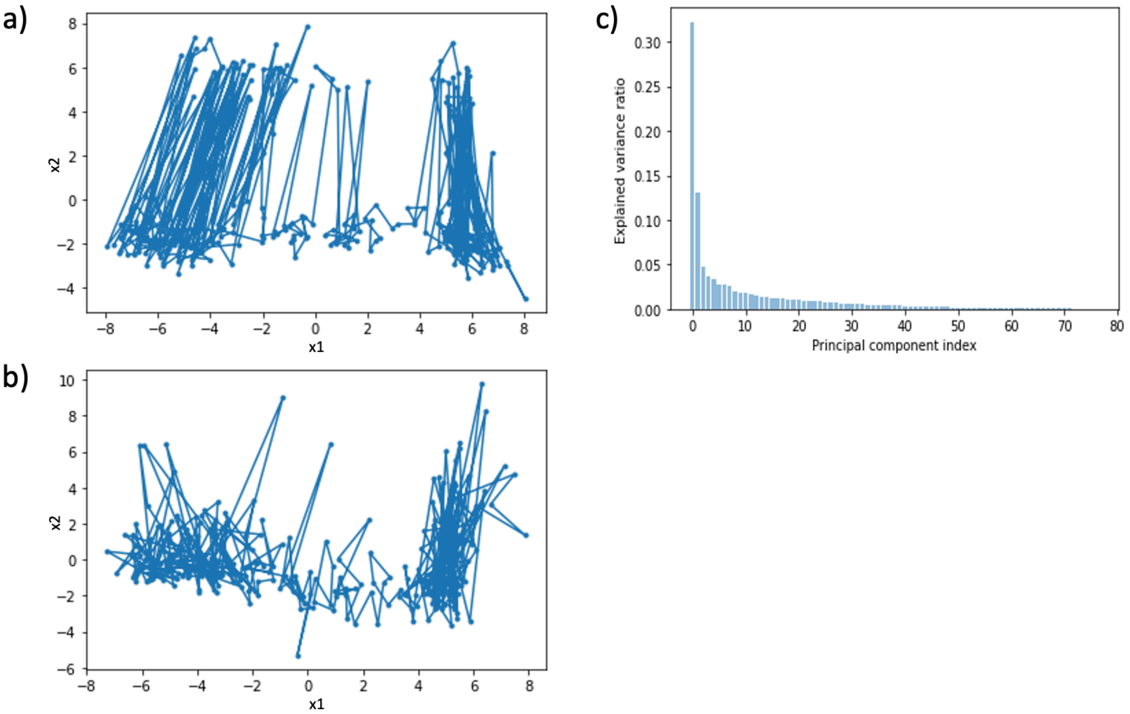


Figure A.14: PCA analysis of the averaged spiking count on trials across session. We computed the average spike count for each neuron and applied PCA. We considered time bins **a.** across the whole trial. Mean activity per trial varies a lot depending on the trial type, mostly on the second principal component. And time bins **b.** on the late ITI only. It is more stable from trial to trial. We see less variation on the second principal component. For both, most of the variation is on the first component, and it drifts across the whole recording. **c.** The explained variance for each component shows that most of the variation is explained by the two first principal components. Overall, if variation on the second principal dimension in **a.** is explained by the structure of the paradigm, variation along the first dimension is general over the whole recording and corresponds to drift along time.

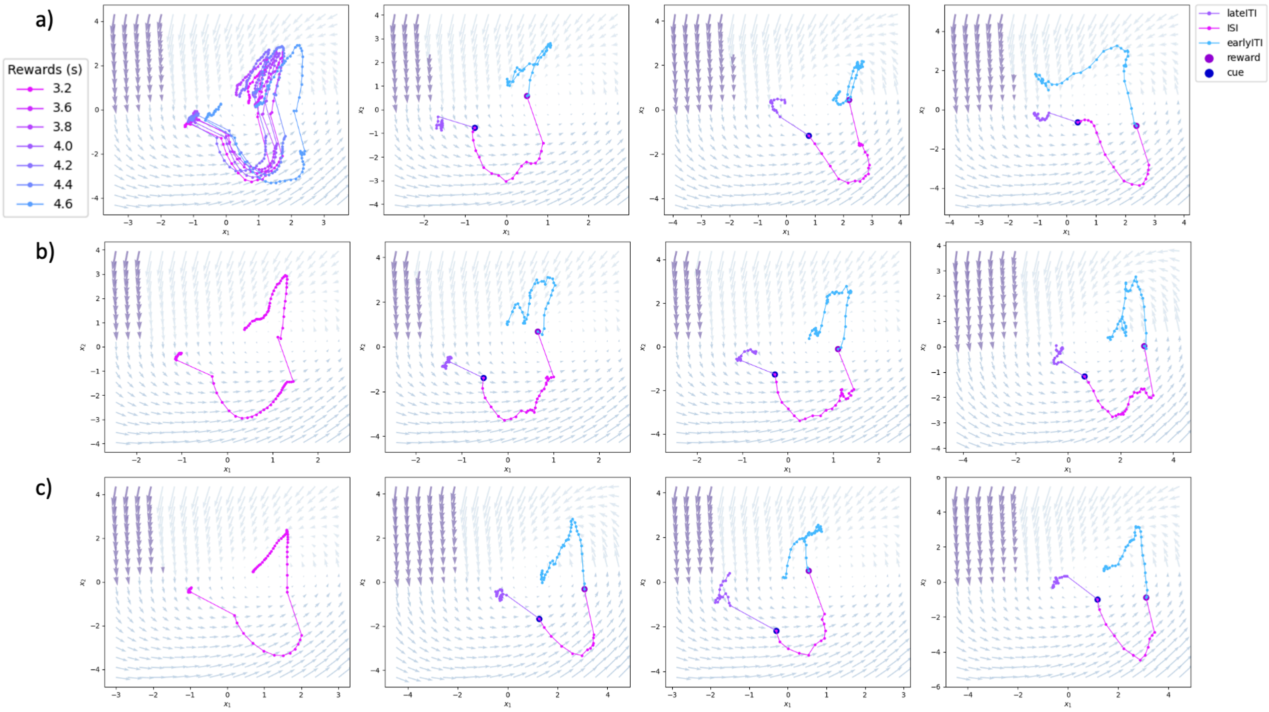


Figure A.15: Latent trajectories per cue type for rSLDS model fitted on OFC neural activity, in which no fixed-point for ISI is visible in the inferred dynamics. a. (Left) Trial-averaged latent trajectory for cue A trials, per reward timing, from 3.2s to 4.8s. Legend corresponds to the time at which the reward happens for each averaged trajectory. **(Right)** Three example individual trials, with difference size of ISI, sampled randomly from the recording. **b. (Left)** Trial-averaged latent trajectory for cue B trials, reward at 3.2s. **(Right)** Three example individual trials, with difference size of ISI, sampled randomly from the recording. **c.** Trial-averaged latent trajectory for cue C trials, for reward at 4.8s. Flow field corresponds to the most likely dynamics in the plane, with each color representing one of the 4 linear systems fitted (one per discrete state) and arrows show the direction and rate of change of activity at each point.

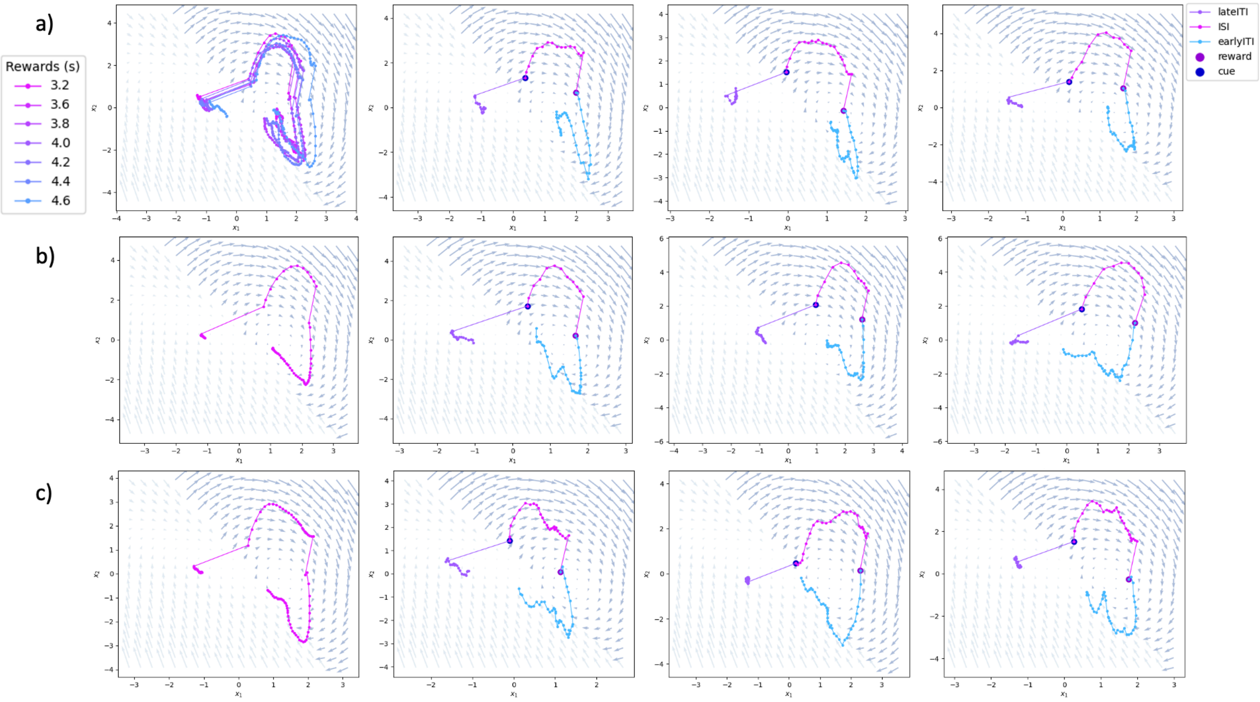


Figure A.16: Latent trajectories per cue type for rSLDS model fitted on OFC neural activity, in which the lat ITI converges to the ISI fixed point. a. (Left) Trial-averaged latent trajectory for cue A trials, per reward timing, from 3.2s to 4.8s. Legend corresponds to the time at which the reward happens for each averaged trajectory. **(Right)** Three example individual trials, with difference size of ISI, sampled randomly from the recording. **b. (Left)** Trial-averaged latent trajectory for cue B trials, reward at 3.2s. **(Right)** Three example individual trials, with difference size of ISI, sampled randomly from the recording. **c.** Trial-averaged latent trajectory for cue C trials, for reward at 4.8s. Flow field corresponds to the most likely dynamics in the plane, with each color representing one of the 4 linear systems fitted (one per discrete state) and arrows show the direction and rate of change of activity at each point.

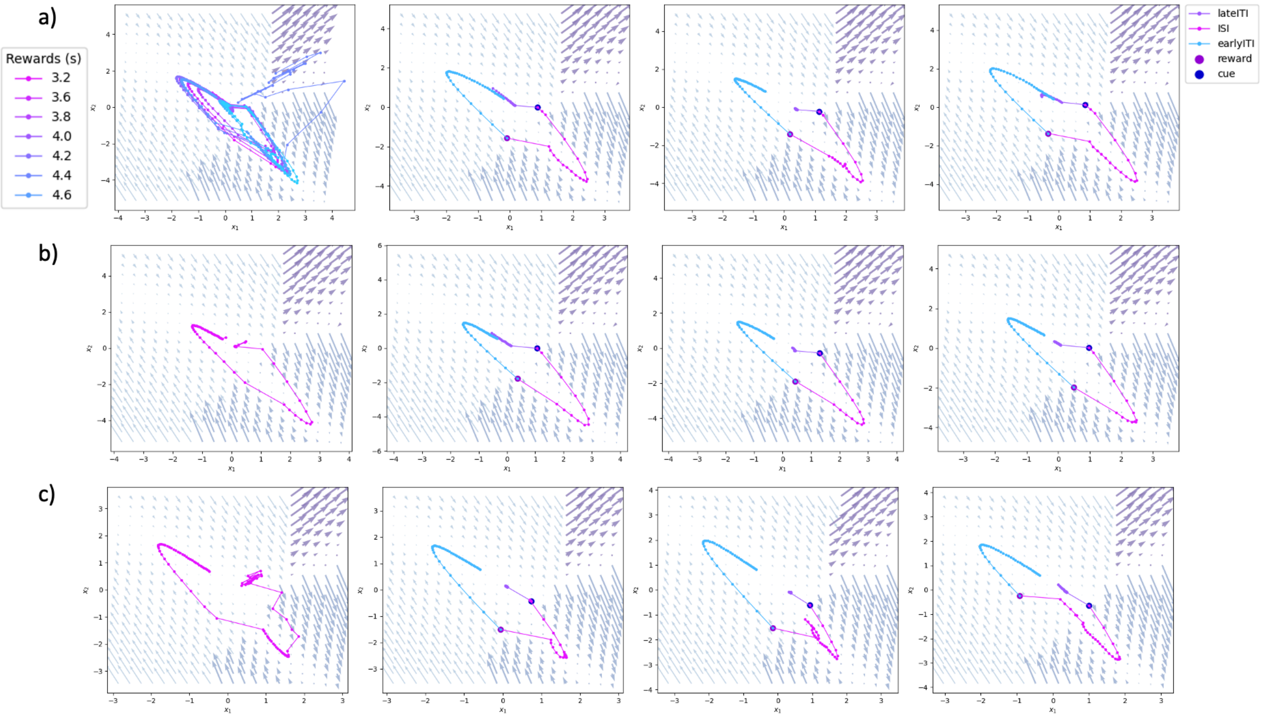


Figure A.17: Latent trajectories per cue type for rSLDS model fitted on OFC neural activity and showing line-attractor dynamics. a. (Left) Trial-averaged latent trajectory for cue A trials, per reward timing, from 3.2s to 4.8s. Legend corresponds to the time at which the reward happens for each averaged trajectory. **(Right)** Three example individual trials, with difference size of ISI, sampled randomly from the recording. **b. (Left)** Trial-averaged latent trajectory for cue B trials, reward at 3.2s. **(Right)** Three example individual trials, with difference size of ISI, sampled randomly from the recording. **c.** Trial-averaged latent trajectory for cue C trials, for reward at 4.8s. Flow field corresponds to the most likely dynamics in the plane, with each color representing one of the 4 linear systems fitted (one per discrete state) and arrows show the direction and rate of change of activity at each point.

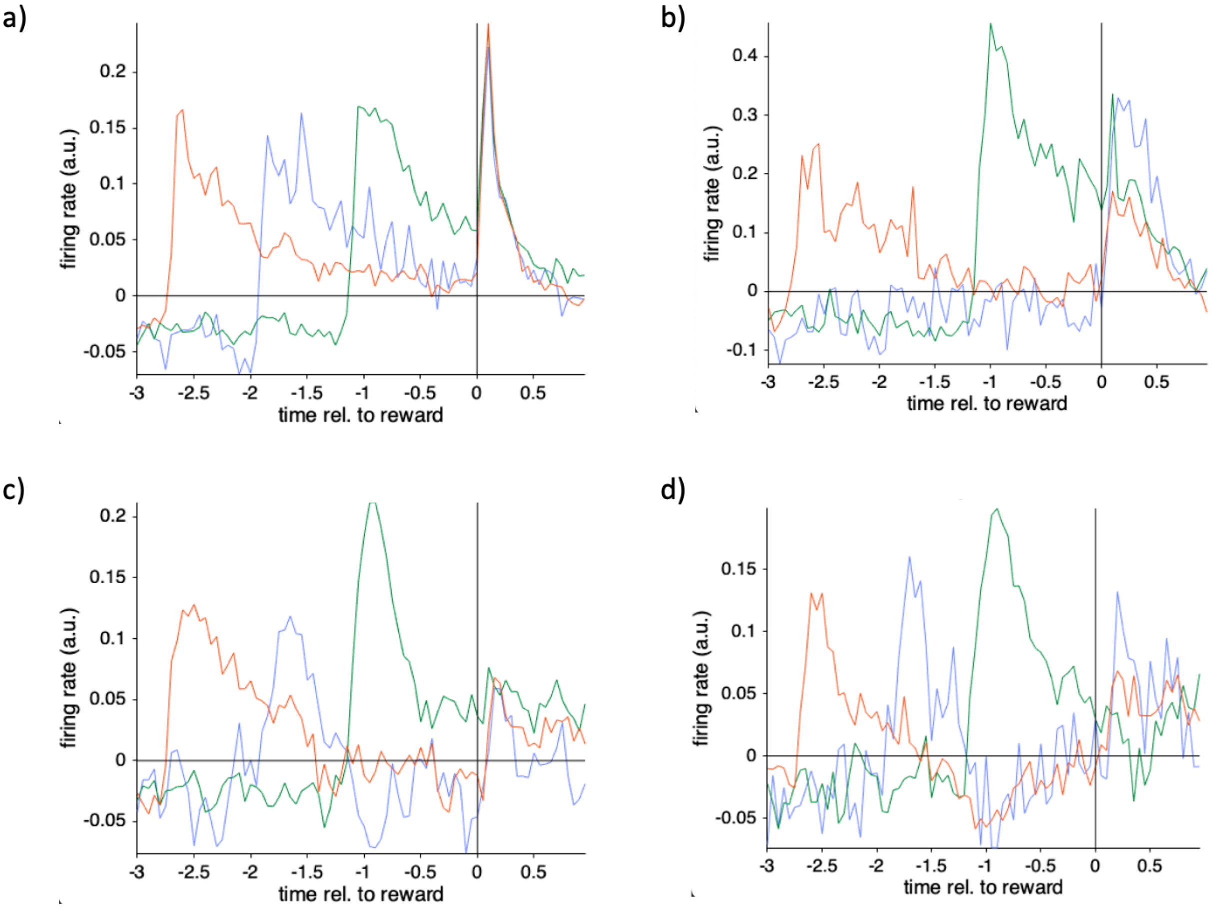


Figure A.18: Trial-averaged neural activity for different sub-regions of the OFC per trial type for task 1. New data were recorded in sub-regions different from the one studied in this project. Trial-averaged neural activity for odor A (blue), odor B (green) and odor C (orange) were compared on a time-scale relative to reward delivery. **a.** Lateral OFC neurons activity recording from the dataset analyzed in the following project. **b.** Lateral OFC neurons activity recording from the newly recorded dataset. Neural activity in trial type A does not display an odor response, further investigation showed that the animal did not seem to have licked neither in the task. **c.** Ventro-lateral OFC neurons activity recording from the newly recorded dataset. Neurons do not seem to be activated upon reward. **d.** Medial OFC neurons activity recording from newly recorded dataset. Neurons do not seem to be activated upon reward. Plots produced by Dr. Jay Hennig.

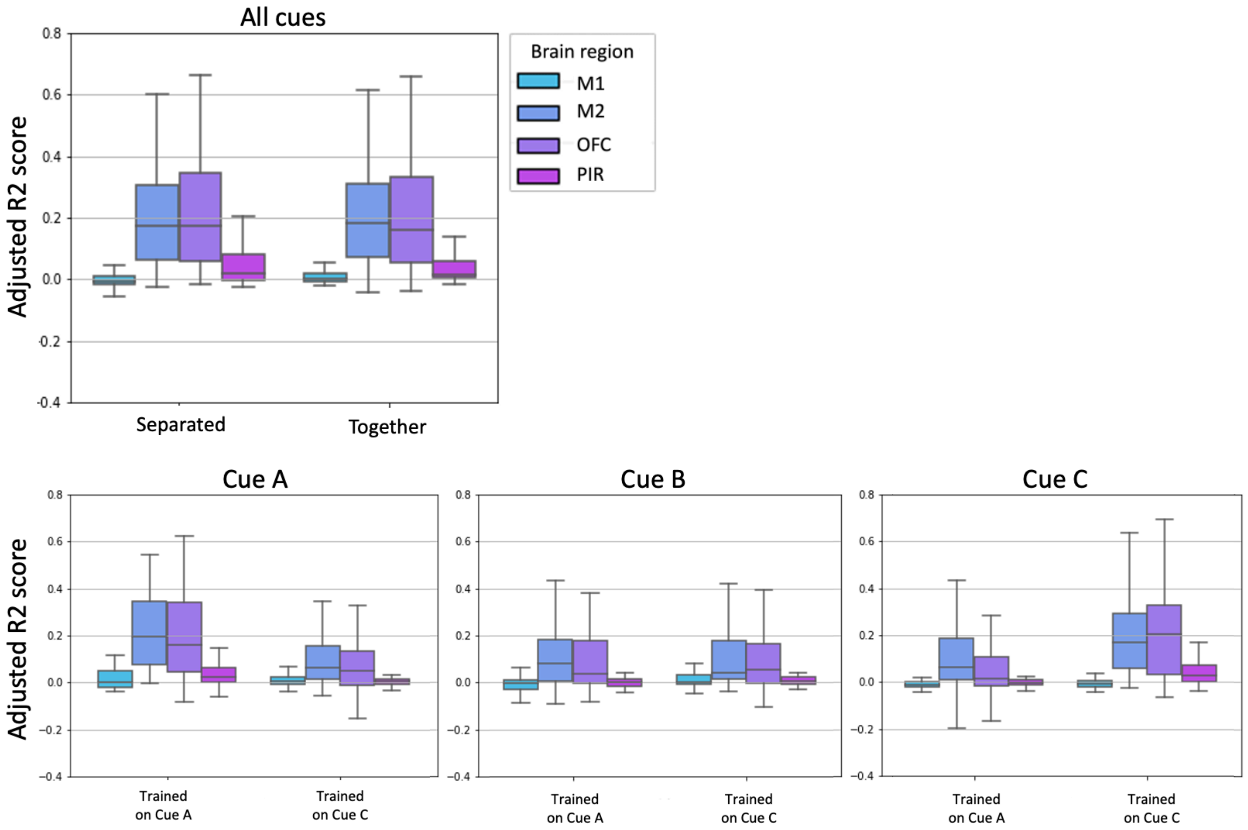


Figure A.19: Distribution of the adjusted R^2-square score across neurons for the LSTSQ linear regression model, trained for each regression matrix and each region on a model with all cue types represented in the same regressors and models only trained on specific cues. Top: Comparison of a model trained on a regression matrix in which each cue type and corresponding reward had its own set of regressors, compared to a model for which the regression matrix had a unique set of regressors for all cues and all rewards. **Bottom:** From left to right, models were trained with all regressors, on only one cue type, either Cue A trials or Cue C trials and adjusted R^2-square scores for individual trial types are displayed, from left to right, Cue A, Cue B and Cue C. Middle line of the box plots is the median score across neurons in the region. We do not display the outliers. Legend holds for all plots.

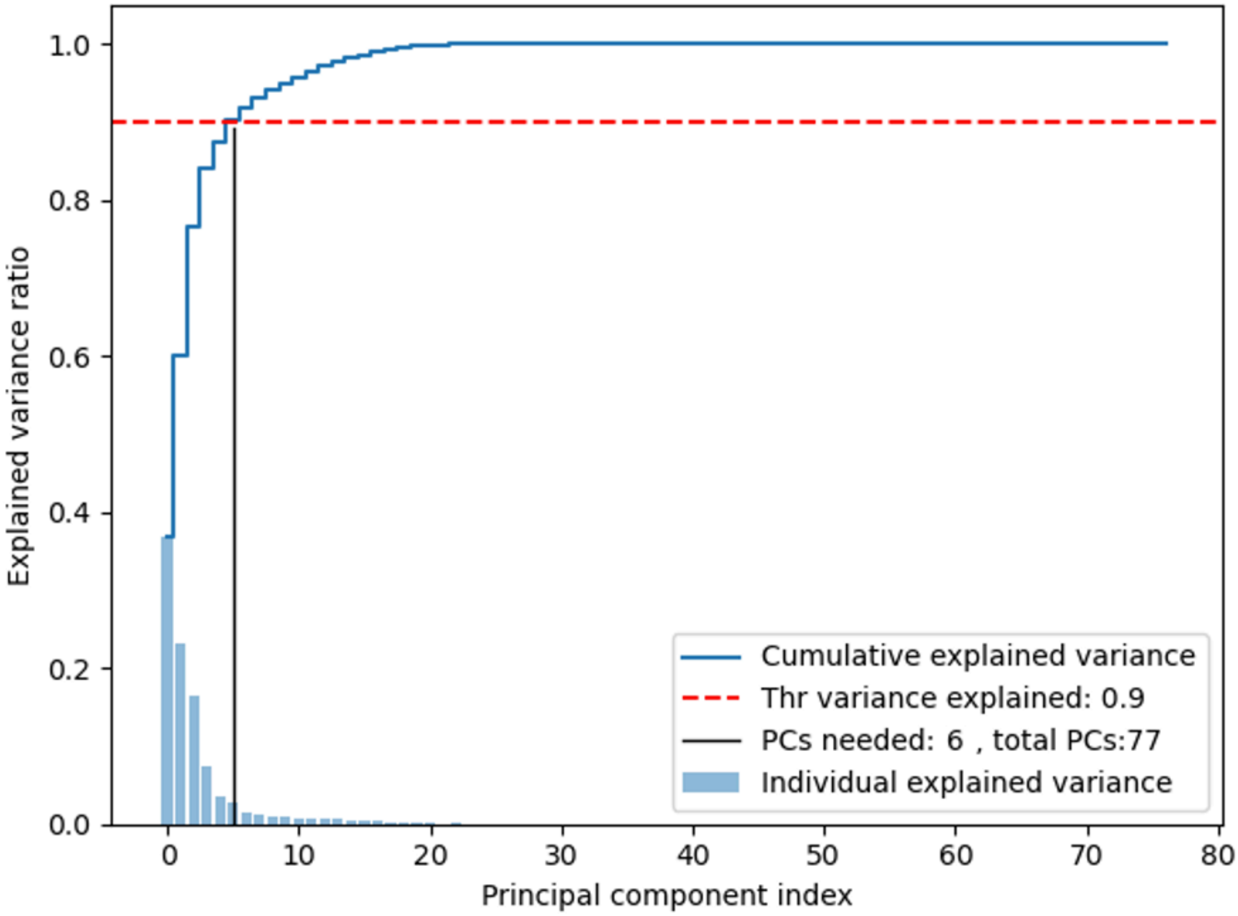


Figure A.20: Cumulative variance of the OFC neural activity explained along the sorted principal factors of a factor analysis. Factor analysis model applied to the neural data for all cues together. Previous work takes the minimum number of principal factors to explain 90% of the variance in the data as the number of latent spaces for the rSLDS model fitting [17].