# What’s going on?

Inference and neural representations of the current situation
and the underlying causal structure of the world

Stephanie C.Y. Chan

A dissertation presented to the faculty of Princeton University
in candidacy for the degree of Doctor of Philosophy

Recommended for acceptance
by the Princeton Neuroscience Institute
Advisers: Yael Niv, Kenneth A. Norman

May 2016

# Abstract

Taken at face value, the world is complicated and confusing. When operating in such complexity, we are greatly advantaged by our ability to infer the *underlying structure* of the world – that is, the relationships between our observations and the underlying *latent causes* that generate them. At any given time, inferring the latent causes that are currently active – i.e., the *current situation* – allows us to execute the most appropriate actions and cognitive processes. In theories of episodic memory, this definition of the current situation is related to the cognitive constructs of “schemas” and “context”. In reinforcement learning and decision-making, representations of the current situation are called the “state”. In this work, I begin to uncover the computations and neural mechanisms that underlie our inference of the causal structure of the world, including inferences of the current situation, and also how the inferred situation affects decision-making and memory. Throughout this work, the overlapping brain areas of ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC) play a prominent role in the neural circuits that perform this inference.

In the first experiment, I show that overall levels of activity in the OFC are related to learning about one type of causal structure – transitions between states of the world. In the second experiment, I present evidence that the OFC represents a belief distribution (a posterior probability distribution) over the underlying situation. In the third experiment, I present evidence that, in accordance with current theories of episodic memory and temporal context, memories seem to be organized according to information in the brain about the semantics of recent experience, which may serve as a heuristic proxy for the current situation.

# Acknowledgements

I have been incredibly lucky to have the support of two amazing advisors, Ken and Yael, who are both wonderful in more ways than I can count. I am certain that I will continue to look up to them as examples and role models, no matter where I end up in life. I am also so grateful to my friends and colleagues at Princeton, who have made it both an incredibly stimulating and incredibly warm environment. Last but not least, I want to express my gratitude towards my family and especially to my parents, who are so wise and who have given me and my siblings more than anyone else in the world.

## Table of Contents

# 1 Organization of the thesis

This thesis presents the results of three empirical investigations into how we make inferences about the causal structure of the world, especially in relation to representations of the current situation, and how we use that information in memory and decision making.

In Chapter 2, I will briefly introduce some background that will be useful for appreciating this work. This introduction includes existing ideas about how the brain represents the current situation and how it affects cognitive processes like memory and decision-making (specifically, I describe the ideas of “state”, “schemas”, and “context”, from theories of reinforcement learning and theories of episodic memory). The introduction also includes background about some of the primary experimental and analytical methods used in this work (relating to multivariate analysis of fMRI data).

In Chapter 3, I present experimental findings showing the involvement of the orbitofrontal cortex in learning about one type of structure in the world – probabilities of transition between different states.

In Chapter 4, I investigate inference and neural representations of the current situation, this time in an environment where the current situation is not directly observable; instead, the situation is probabilistically related to subjects’ observations. I show that humans use an inference strategy that approximates Bayesian inference of a posterior probability distribution over the underlying situation, and that the orbitofrontal cortex represents this probability distribution over the underlying situation.

In Chapter 5, I provide experimental evidence for a key, as yet untested, assertion in current theories of episodic memory – that memories are organized according to the semantics of recent experience. I show that memories are more likely to be recalled together if they were encoded at a time when the brain showed similar information about the semantics of recent experience. Since the semantics of recent experience can serve as a heuristic stand-in for the current situation, it makes sense for memories to be organized according to these semantics, so that memories from a given situation are retrieved when that situation is encountered again.

In Chapter 6, I discuss the experimental results in relation to one another, and unanswered questions that remain in investigating cognitive and neural representations of the current situation, and their role in decision-making and memory.

# 2 Background

## 2.1 Representations of the current situation – existing ideas and open questions

This dissertation draws on existing ideas relating to how we represent and infer the current situation, both algorithmically and neurally. Below, I first introduce the ideas of “state” (from reinforcement learning), and “schemas” and “context” (from theories of memory). I also introduce some of the existing evidence that the orbitofrontal cortex (OFC) is involved in processing these signals. I also summarize the open questions that will be addressed by the experimental results presented in the later chapters of this thesis.

### 2.1.1 State

In reinforcement learning (e.g. Sutton and Barto, 1998), an agent learns to assign values to different environmental “states”, based on rewards and punishments received. The learned values estimate the total future reward that is expected after entering that state (including but not limited to the rewards received at that state). These values can be learned using error-driven learning algorithms (e.g. Rescorla et al, 1972), which includes temporal-difference learning methods (e.g. Sutton 1988). Temporal difference learning algorithms compare predictions with new information about values, and update cached values based on the discrepancy (the “prediction error”). Once the value for each state is learned, the agent can use this information to make decisions about what actions to take in what state in order to maximize total expected reward, by taking actions that lead to states with high value.[1]

How should states be defined? If you are sitting in a restaurant, for example, you might define your state according to all the details of all your observations, including the color of the tablecloth, the position of your knife with respect to the plate, and so on. However, if you are trying to hold a conversation with your dinner partner, this information is not useful and does not need to be included in the state representation in order for you to perform well at the conversation. Instead, you might even require information that is not directly observable and which needs to be inferred from observations or held in memory, such as the current topic of conversation or the topics that have already been discussed. Thus, some state representations are more useful and natural than others, and they may include information that is not directly observable.

The recursive nature of reinforcement learning algorithms requires states to be defined so that the dynamics of the task, at any given time, are completely determined by the current state (and do not require information about previous states). This is the Markov property. In some cases, as in the

[1] To be precise, there exist some policy-gradient algorithms for reinforcement learning that do not estimate values at all, though most do. These algorithms nonetheless learn a policy that is defined over states, so that proper definition of states is still key, including the Markovian property that we discuss below.

restaurant conversation example, the dynamics of the task are dependent on information that is not directly observable (a partially observable environment) – that is, the observations are not Markov with respect to the task, and the machinery of reinforcement learning cannot be applied to state representations that contain only the observations. However, it turns out that even in these types of tasks, reinforcement learning algorithms can be applied to a *belief distribution over states* (a “belief state”), which is Markov (Kaelbling et al 1998).

There is by now a wealth of evidence that the brain implements some form of temporal-difference reinforcement learning. Prediction errors appear to be represented by dopamine neurons, while learned values are to be represented in the striatum (e.g. Montague et al, 1996; Schultz et al, 1997; O’Doherty et al, 2004; Lau and Glimcher, 2008). More recently, it has been suggested that the OFC is involved in representations of state (Wilson et al, 2015). I test this hypothesis in Chapter 4, additionally showing that the representation in OFC takes the form of a belief distribution over possible states, as would be necessary for the common case of a partially observable environment. This result relates also to the implicated representation of schemas in OFC, discussed in the next section.

In certain cases, expected values learned incrementally through trial-and-error (so called “model-free” values) cannot support optimal decision-making, for instance because the agent does not yet sufficient experience with the environment to estimate correct values. If, for example, the environment has not been well explored or there is a change in the environment, it is often useful to *simulate* sequences of states and/or actions, so as to supplement the meager amount of actual experience. This type of decision-making is called “planning”, “goal-directed”, or “model-based” decision making, in contrast with “model-free” decision making. For example, when encountering a new type of food, we may try to mentally simulate the experience of eating it before making a decision on whether to eat it (model-based); with very familiar types of food, we are more likely to have a pre-computed approach or avoidance response (model-free). These two types of decision making appear to be implemented in neural circuits that are parallel but not completely overlapping (Daw et al, 2005). For model-based decision making, it is necessary to learn about the probability of transitions between states, sometimes called the “transition matrix” for the environment. The OFC has long been implicated in model-based decision making. In Chapter 3, I investigate the role of OFC in learning these transitions between states, in addition to its presumed role in representing the states themselves.

### 2.1.2 Schemas and Context

In theories of episodic memory, it is believed that we organize our memories according to an inferred “schema” that specifies the gist of a situation, and provides previously learned associations that a new memory can be incorporated into (Tse et al, 2007; Hupbach et al, 2008). Recent evidence has pointed to the medial prefrontal cortex in processing or representing schemas (for reviews, see Schlichting and Preston, 2015; van Kesteren et al, 2012; Ranganath and Ritchey, 2012). For example, Tse et al (2011) showed evidence that activation of rat medial prefrontal cortex (mPFC) was highest immediately after memory encoding that should involve incorporation of new information into existing schemas, and also that transient inactivation of that area blocked retrieval of consolidated memories. In humans, Ezzyat and Davachi (2011) showed that greater activation of ventromedial PFC (vmPFC; a subregion of mPFC) in humans during memory encoding was correlated with how strongly those memories were associated with other memories in the same "event", consistent with the idea that vmPFC is involved in schemas that are bound to memories. Note that vmPFC and OFC (implicated in state representations for reinforcement learning, as described in the previous section) have varying anatomical definitions, but are often construed to be overlapping.

Even while the field has made gains in uncovering the neural basis of schema representation, it remains to be specified concretely what constitutes a schema. In Chapter 4, I investigate the hypothesis that schemas are inferred using Bayesian latent cause inference, and that they are represented in vmPFC/OFC as a probability distribution over the possible "latent causes". Bayesian latent cause inference allows us to make inferences about what underlying causes are generating our observations at any given time, with the statistically optimal inference involving the computation of a posterior probability distribution over latent causes. Viewing the world as structured according to latent causes may serve as a general organizing principle for our memories. Our results from Chapter 4 support previous ideas that vmPFC/OFC represents inferred situation, and, moreover, suggest that this representation does indeed take the form of a posterior probability distribution.

A related idea in theories of episodic memory is the idea of “context”. It is believed that all memories are encoded with information about their context, which may include information about time, space, and emotional state. Context is thought to organize memories and to act as a retrieval cue, so that memories are activated when a similar context is reinstated. A key untested claim in current theories of episodic memory (e.g. Howard and Kahana, 2002; Polyn et al, 2009) is that memories are organized by the semantics of recent experience; i.e. memories are more likely to be reinstated when the experiences immediately preceding the memory are semantically similar to experiences that precede the current timepoint, and also memories preceded by semantically similar experiences tend to be recalled together. In Chapter 5, I test these predictions and provide evidence for this claim.

## 2.2 Methods used for neural analyses – multivariate analyses and machine learning with fMRI

In recent years, human fMRI studies have benefited greatly from the development of multivariate analyses and the application of machine learning to fMRI data. These methods examine *patterns* of activity across voxels in the brain, without requiring specific predictions about what exactly those patterns of activity should look like.

In the following work, I use two main classes of multivariate analysis: (1) the application of classification methods from machine learning (e.g. Norman et al, 2006; Lewis-Peacock and Norman, 2015), and (2) "pattern similarity" analysis, which examines the similarity structure of patterns of neural activity, and identifies brain regions where the similarity structure of neural patterns matches the similarity structure predicted by a particular cognitive model (e.g. Kriegeskorte et al, 2008).

Classification analyses apply methods from machine learning to classify data according to given class labels. Common methods include logistic regression and support vector machines. If patterns of activity from an area of the brain can be successfully classified according to the desired class labels, that indicates encoding of class information in the brain area.[3] I apply these methods in Chapters 3 and 5.

Pattern similarity analyses can be performed when we have a model that makes predictions about the similarity structure of cognitive states – i.e., groups of time points in the experiment where we expect cognitive states to

[3] However, care should be taken not to overinterpret successful classification in a brain area; for example, it does not imply that the brain area itself processes a similar type of classification.

be more or less similar. For example, if we were trying to identify a brain area that represents colors, we could measure brain activity in response to several different colors, and try to identify an area whose patterns of activity look more similar when the colors are more similar. I apply pattern similarity analyses in Chapter 3.

Without a specific *a priori* hypothesis about a region of interest in the brain, “searchlight” methods allow application of multivariate analyses to every part of the brain, applying the analysis iteratively to small neighborhoods of voxels across the brain, one neighborhood at a time. I employ searchlight analyses in Chapters 3 and 5.

We now turn to the three empirical studies, in Chapters 3, 4, and 5.

## 2.3 References

Daw ND, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience. 8:1704–1711.

Ezzyat Y, Davachi L. 2011. What Constitutes an Episode in Episodic Memory? Psychological Science. 22:243–252.

Howard MW, Kahana MJ. 2002. A Distributed Representation of Temporal Context. Journal of Mathematical Psychology. 46:269–299.

Hupbach A, Hardt O, Gomez R, Nadel L. 2008. The dynamics of memory: Context-dependent updating. Learn Mem. 15:574–579.

Kaelbling LP, Littman ML, Cassandra AR. 1998. Planning and acting in partially observable stochastic domains. Artificial intelligence. 101:99–134.

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in systems neuroscience. 2:4.

Lewis-Peacock JA, Norman KA. 2015. Multi-Voxel Pattern Analysis of fMRI Data. In: Gazzaniga M,, Mangun R, editors. Cognitive Neurosciences V. Cambridge, MA: MIT Press. p. 911–920.

Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 16:1936–1947.

Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences. 10:424–430.

Polyn SM, Norman KA, Kahana MJ. 2009. A context maintenance and retrieval model of organizational processes in free recall. Psychol Rev. 116:129–156.

Rescorla RA, Wagner AR. 1972. A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In: Black AH,, Prokasy WF, editors. Classical conditioning II: current research and theory. New York: Appleton-Century-Crofts. p. 64–99.

Schultz W, Dayan P, Montague PR. 1997. A Neural Substrate of Prediction and Reward. Science. 275:1593–1599.

Sutton RS. 1988. Learning to predict by the methods of temporal differences. Machine learning 3:9–44.

Sutton RS, Barto AG. 1998. *Reinforcement Learning: An Introduction*. Cambridge, Mass: A Bradford Book.

Tse D, Langston RF, Kakeyama M, Bethus I, Spooner PA, Wood ER, Witter MP, Morris RGM. 2007. Schemas and Memory Consolidation. Science. 316:76–82.

Tse D, Takeuchi T, Kakeyama M, Kajii Y, Okuno H, Tohyama C, Bito H, Morris RGM. 2011. Schema-Dependent Gene Activation and Memory Encoding in Neocortex. Science. 333:891–895.

Wilson RC, Takahashi YK, Schoenbaum G, Niv Y. 2014. Orbitofrontal Cortex as a Cognitive Map of Task Space. Neuron. 81:267–279.

# 3 Learning to predict state transitions and the orbitofrontal cortex

This work was undertaken in collaboration with Nicolas W. Schuck, Nina Lopatina, and Yael Niv. This work has previously presented at the following conferences: *Society for Neuroeconomics*, Miami, FL (2015); *Workshop on the Neurobiology of Prediction and Surprise*, New Brunswick, NJ (2014); Reinforcement Learning and Decision Making, Princeton, NJ (2013). It is currently in preparation for submission as a journal article.

## 3.1 Introduction

To flexibly plan for the future, we must be able to predict which states of the world lead to which (the “transition structure” of the world). For example, if we know that drinking coffee makes us perky but warm milk makes us sleepy, we can make plans about what beverage to drink at different times of the day. This type of planning has been termed “model-based decision making” (Daw et al, 2005), for which the orbitofrontal cortex (OFC) has been shown to be particularly important (Baxter et al, 2000; Izquierdo et al, 2004; Valentin et al, 2007; De Wit et al, 2009; McDannald et al, 2011; Rudebeck et al, 2011; Wilson et al, 2014). However, previous research has concentrated on showing that OFC activity relates to the expected values of future rewards (Schoenbaum et al, 1998; Gottfried et al, 2003; Padoa-Schioppa and Assad, 2006; Hampton et al, 2006; Fellows, 2007; Hare et al, 2008; Wallis and Kennerley, 2011; Monosov and Hikosaka, 2012), a role that does not explain why the OFC is critical specifically for model-based planning. Recently, a study in rodents suggested a different or additional role for the OFC – learning about changes in transition structure (McDannald et al, 2011). Learning state transitions is critical to model-based planning, because one cannot mentally simulate the future result of current actions without an accurate model of how these transitions are likely to unfold in the future. We therefore set out to test whether the OFC might be involved in error-driven learning about state transitions.

Specifically, we hypothesized that the OFC computes or represents a prediction error at the time of unexpected outcomes, which can be used to update an internal model of the transition structure of the world. A previous study implicated several brain areas in computing such “state prediction errors” (Gläscher et al, 2010), but not the OFC. However, the OFC is particularly difficult to image using fMRI due to drop-out and susceptibility artifacts, which lead to low signal-to-noise ratios (Deichmann et al, 2003). Effects in OFC might therefore be difficult to detect in whole brain studies that do not specifically target the OFC as an a priori region of interest. Here we performed a targeted investigation of the OFC, using imaging parameters that maximize signal in this area, and testing our hypotheses specifically in the area. We also tested for effects of OFC activity on behavioral measures of learning transition structure, on a trial by trial basis.

State prediction-error signals should occur upon observing state transitions that are unexpected, and can be used to guide learning so that these transitions are better predicted in the future. These error signals are analogous to – but distinct from – reward prediction errors that are used for learning to associate states with their reward values (e.g., Rescorla and Wagner, 1972; Montague et al, 1996). In particular, reward prediction errors occur upon observing unexpectedly large or small rewards, while state prediction errors should occur even when the reward value is just as expected, e.g. when transitioning to a state that was unexpected but just as valuable as the state that was expected. For example, if you discover that your beer bottle is in fact full of wine, you will experience a state prediction error, even if you dislike beer and wine equally.

In our experiment, subjects performed a task designed to elicit state prediction errors in the absence of reward prediction errors. In this task, black-and-white image cues led probabilistically to different quantities and colors of M&M candies (outcomes). In the critical trials, the number of M&Ms was fully predictable, but their color was not – subjects should therefore experience a state prediction error even though the value of the outcomes was not surprising. Subjects were hungry and were rewarded with actual M&Ms at the end of the experiment. Using fMRI, we investigated activity in the OFC at the time of the outcomes.

If the OFC processes prediction-error signals for learning about state transitions, then we expect that OFC activity should be different for large vs. small state prediction errors (i.e. when an observed state outcome is more vs. less surprising). Furthermore, we expect that greater OFC activity at the time of an outcome should correspond to greater learning about that outcome, and hence greater expectations of that outcome in the future. We found that OFC did show a response at the time of outcomes, and that this response was indeed correlated with learning about the state transitions. Using multivariate pattern analysis (MVPA) on BOLD activity in the OFC at the times of the outcome, we also found that we could successfully decode representation of the outcome – a prerequisite for learning about the correct state transition. However, we found that we could not distinguish OFC responses for more vs. less surprising outcomes, and also OFC activity was correlated with learning, but not in the way we expected. These results suggest that the OFC does indeed contribute to learning, but not via the particular error-driven learning algorithm that we originally hypothesized.

## 3.2 Methods

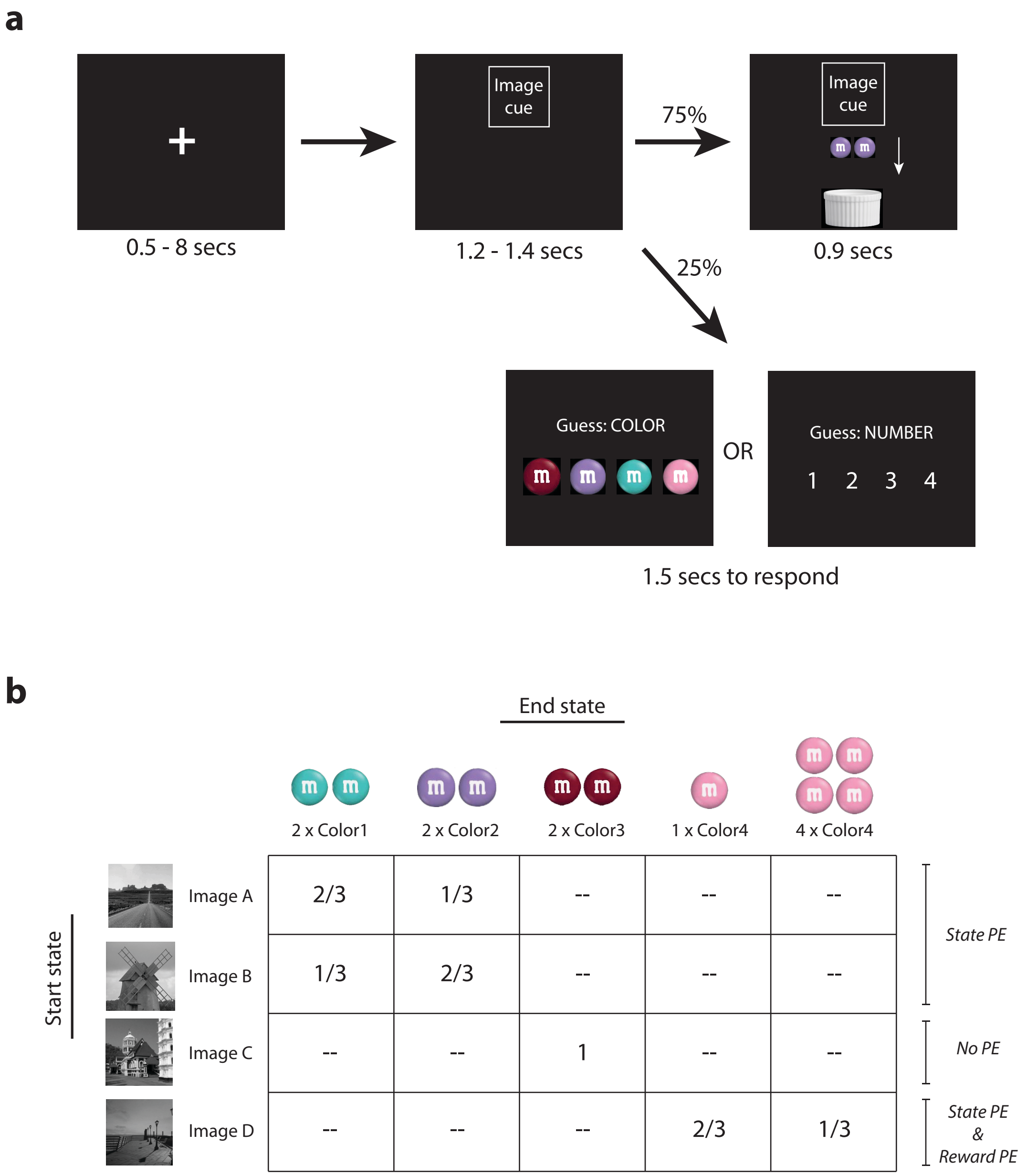
**Participants.**

Twenty-four volunteers from the Princeton University community participated in exchange for monetary compensation ($20 per hour + up to $10 performance-related bonus). All participants were right-handed (14 female, age range 18-34 years) and stated that they liked M&Ms. Informed written consent was obtained from all participants, and the study protocol was approved by the Institutional Review Board for Human Subjects at Princeton University.

## Experimental design

Each trial began with 0.5 - 8 seconds of fixation (exponentially distributed, mean 2.4 s). Then one of four black-and-white image cues depicting outdoor scenes appeared for 1.2 s (see Fig 1a). On 75% of the trials, this was followed by the opening of a box around the image (0.2 s). Then, a set of M&Ms appeared below the image and fell into a bowl, over the course of 0.9 s. As the M&Ms fell into the bowl, one clinking sound was emitted for each M&M in the set. A tally at the bottom of the screen indicated the total number of M&Ms received in the experiment so far, for each of the four possible colors (not shown in Fig 1a).

Each of the four image cues was associated with different numbers and colors of M&Ms according to a predetermined schedule of reinforcement (Fig 1b). **Image A and Image B** were designed to elicit state prediction errors throughout the experiment due to a probabilistic schedule of M&M color, but not reward prediction errors, because they always dropped exactly 2 M&Ms. **Image C**, in contrast, was associated with 2 M&Ms of a fixed color, thus eliciting no prediction errors once the contingencies had been learned. Finally, **Image D** was designed to elicit reward prediction errors—it was associated with a fixed color of either 1 or 4 M&Ms on different trials (as with the other image cues, Image D led to 2 M&Ms on average). For each subject, the images and M&M colors were assigned randomly from a pool of 20 images and 5 M&M colors.



| Start state | 2 x Color1 | 2 x Color2 | 2 x Color3 | 1 x Color4 | 4 x Color4 | |
|---|---|---|---|---|---|---|
| Image A | 2/3 | 1/3 | -- | -- | -- | *State PE* |
| Image B | 1/3 | 2/3 | -- | -- | -- | |
| Image C | -- | -- | 1 | -- | -- | *No PE* |
| Image D | -- | -- | -- | 2/3 | 1/3 | *State PE & Reward PE* |

**Figure 1. (a) Experimental design.** Trials began with fixation. Then, one of four image cues ("start states") appeared. On most trials, the box around the image opened, and a number of colored M&Ms ("end states") dropped from the image, clinking as they fell into a bowl. On the randomly interspersed "guess" trials, the image cue was instead followed by a prompt to guess (within 1.5 seconds) either the color or number of M&Ms that would have fallen on that trial. **(b) Cue-outcome contingencies for each of the four images (transition matrix for the experiment).** Numbers in table indicate probability of each end state (M&M outcome) given each start state (image cue). PE = prediction error. Larger state prediction errors are expected for rarer outcomes (smaller transition probabilities). Images and M&M colors were assigned randomly for each subject. Our analyses focused on Image A and Image B trials, which were designed to elicit state prediction errors in the absence of reward prediction errors.

Subjects earned 1 real M&M of a given color for every 17 “virtual” M&Ms that they received in the task. Subjects were requested to refrain from eating or drinking (except water) for at least 3 hours prior to the experiment, so that the M&Ms would be especially rewarding. Non-standard M&M colors were chosen to circumvent pre-existing preferences for specific M&M colors, and to achieve perceptually distinct outcomes that are of equal value. (Note also that our analyses of state prediction error always combined Image A and Image B trials, so that any potential value differences between the two colors would cancel out.) In a post-experiment questionnaire, subjects rated the appeal of the M&Ms on a scale from 1 (not appealing at all) to 5 (very appealing). The mean rating was 3.8 ± 0.2.

On “guess trials” (25% of all trials, pseudorandomly distributed), the appearance of the black-and-white image cue was followed by a prompt reading “Guess: COLOR” or “Guess: NUMBER”. At the appearance of the prompt, the image cue disappeared. Subjects were given 1.5 s to guess what color/number of M&Ms *would have* fallen on that trial. Subjects received 10¢ for every question correctly answered. The purpose of the guess trials was to encourage subjects to pay attention to the image cue and to actively make a prediction of the upcoming M&M outcome on *every* trial – because the allowed response time was so short, subjects had to prepare an answer upon viewing the image cue in case a guess prompt followed.

Subjects performed 72 training trials outside of the scanner, to familiarize themselves with the task and to learn the stimulus-outcome contingencies. During training, subjects received and ate the M&Ms they earned (approximately 7 M&M candies). They were then informed that future M&Ms they earned would be given to them after the ensuing scanning session, and they performed another 420 trials in the MRI scanner. At the end of the experiment, subjects received all M&Ms earned while in the scanner. The 420 trials were evenly distributed between the four image cues, with trial order pseudorandomized so that the total number of M&Ms collected increased at the same rate for every color. The experiment was divided into 5 scan sessions of approximately 10 minutes each.

**Behavioral measures**

We evaluated three types of behavioral measures, computed separately for each subject and for each prediction trial type (image cue type x number/color prediction): (1) overall performance over the course of the experiment; (2) change in performance over the course of the experiment (3) sensitivity to the most recent outcome (a proxy for learning rate).

For (1), we computed the fraction of responses that were optimal (i.e. for which the subject selected the common outcome). We excluded the training session for this measure (only including responses from the scan sessions), although the results are very similar when we also include the training session.

For (2), we computed the difference in performance from the beginning to the end of the experiment, given as (fraction of optimal responses in the last scan session) – (fraction of optimal responses in the training session).

For (3), we computed the probability of predicting the common outcome after observing the *common* outcome on the previous trial with the same image cue, and compared with the probability of predicting the common outcome after observing the *uncommon* outcome on the previous trial with the same image cue. The difference between these two quantities serves as a proxy for learning rate – subjects with high learning rate would be more sensitive to the most recent outcome, and would show a larger difference between the two quantities.

### fMRI acquisition

Functional brain images were acquired using a 3T MRI scanner (Skyra; Siemens Erlangen, Germany), and were preprocessed using FSL (http://fsl.fmrib.ox.ac.uk/fsl/). An echoplanar imaging sequence was used to acquire 40 slices of 2mm thickness with a 1-mm gap (repetition time (TR) = 2.4s, echo time (TE) = 27ms, flip angle = 71°, field of view = 196 mm, phase encoding direction = anterior to posterior). We optimized our fMRI sequence for OFC signal acquisition by including a gap between slices, using shimming and fieldmap unwarping, and tilting the slices by approximately 30° from the axial plane towards a coronal orientation (Deichmann et al, 2003). Fieldmaps consisted of forty 3-mm slices, centered at the centers of the echoplanar slices, with TR = 500ms, TE1 = 3.99 ms, TE2 = 6.45ms, field of view = 196mm. At the end of the 5 functional scanning sessions, an MPRAGE anatomical scan was acquired, consisting of 176 1-mm axial slices, TR = 2.3s, TE = 3.08 ms, flip angle = 9°, and field of view = 256mm.

### Preprocessing

All functional images were preprocessed using low pass filtering (filter at 1/100 Hz), motion correction (six-parameter rigid body transformation), correction for B0 magnetic inhomogeneities (fieldmap unwarping), spatial smoothing (Gaussian kernel with full width at half maximum of 5mm), and co-registration of functional and structural scans. For GLM results, we additionally performed spatial normalization of subject-level results to match a template in MNI space (12-parameter affine transformation).

### Functional parcellation of orbitofrontal cortex

Regions of interest for the orbital frontal cortex were obtained from Kahnt et al, who used k-means clustering of functional connectivity patterns to parcellate OFC into subregions (Kahnt et al, 2012). We used the parcellation of OFC into two clusters, which corresponded with medial-lateral subdivisions of OFC found in studies of cytoarchitectonic structure and of intra-regional anatomical connectivity (Carmichael and Price, 1996; Ongür and Price, 2000).

## Obtaining mean percent signal change at M&M outcomes

Using the FSL toolbox (http://fsl.fmrib.ox.ac.uk/fsl/), we performed a GLM with the following regressors: one regressor for the onsets of each type of image cue (A, B, C, D); one regressor for the onsets of the M&M outcomes for Image C; one regressor each for the onsets of the uncommon outcomes of Images A, B, and D; one regressor each for the onsets of the common outcomes of image cues A, B, and D; and one parametric regressor for the clinks of the M&Ms into the bowl (1, 2, or 4 clinks). These regressors were convolved with a standard hemodynamic response function, and combined with 6 motion regressors.

Regressor weights for each voxel and each scan session were converted to percent signal change by multiplying by the appropriate scale factor for events of length 0.1 sec convolved with the standard double-gamma hemodynamic response function, and then dividing by the mean of the voxel's timecourse for that scan session. These per-scan numbers were averaged across scans for each participant. To obtain the percent signal change for a region of interest, the percent signal change was averaged across all voxels in the region of interest.

## Obtaining trial-by-trial estimates of percent signal change at M&M outcomes

To obtain trial-by-trial estimates of percent signal change (PSC) in an ROI at each M&M outcome, we fit a separate GLM for each trial. This GLM was identical to the one used for estimating mean PSC (above), except that the regressor for the condition of the trial of interest was split into one regressor modeling the onset just for the trial of interest, and a second regressor modeling the onsets of all other trials in that condition (Mumford et al, 2012). These GLMs were fitted on data that was preprocessed in FSL, but (for computational reasons) the GLMs themselves were fitted in MATLAB.

### MVPA classification

The purpose of our MVPA analyses was to see whether activity in OFC at the time of the M&M outcomes contained information about the start state and end state (stimulus and outcome) for each transition.

Given our rapid event-related design, we first used a GLM to deconvolve neighboring events, regress out motion artifacts, and to de-noise examples through averaging (Mumford et al, 2012). The GLM included, for each half of each scan session, regressors modeling the appearance of the M&Ms for each of four trial types (Image A followed by M&M Color 1, Image A followed by M&M Color 2, Image B followed by M&M Color 1, Image B followed by M&M Color 2), comprising of 8 regressors per run. These were convolved with a canonical hemodynamic response function. In addition, for each scan session we modeled head motion using six motion regressors and the mean activity using an intercept regressor. We estimated this GLM on each subject's smoothed, motion-corrected fMRI data using the FSL toolbox (http://fsl.fmrib.ox.ac.uk/fsl/).

We used the resulting patterns of voxel-wise regressor weights for the four trial types (two regressor weights per run and trial type; z-scored) as training and testing examples for a support vector machine (SVM) classification algorithm with a linear kernel (nu-SVM, as implemented in LIBSVM, Chang and Lin, 2011), under a leave-one-session-out cross validation scheme using the Princeton MVPA Toolbox (https://code.google.com/p/princeton-mvpa-toolbox). We used a standard cost (nu) parameter of 1 for the SVM (the results did not vary much with this parameter).

To classify start state, we classified training and testing examples according to the image cue (Image A or Image B). To classify end state, we classified training and testing examples according to the M&M color (Color 1 or Color 2).

## 3.3 Results

**Overall behavioral performance on prediction task**

For the prediction task, the optimal strategy was to predict the most common outcome on every trial. Overall, subjects predicted the most common outcome 77 ± 2% of the time. The 23% non-optimal guesses may have resulted from a combination of probability matching (for probabilistic transitions, Vulkan, 2000; Erev and Barron, 2000), imperfect knowledge of transition probabilities, and noise. Fig 2a shows subject performance on each trial type. For the probabilistic trial types, participants were, on average, close to true probability matching, for which we would expect 67% optimal responses. Performance on color prediction for Image A was significantly greater than for Image B (p = 0.005), likely because Image A was presented first during training.

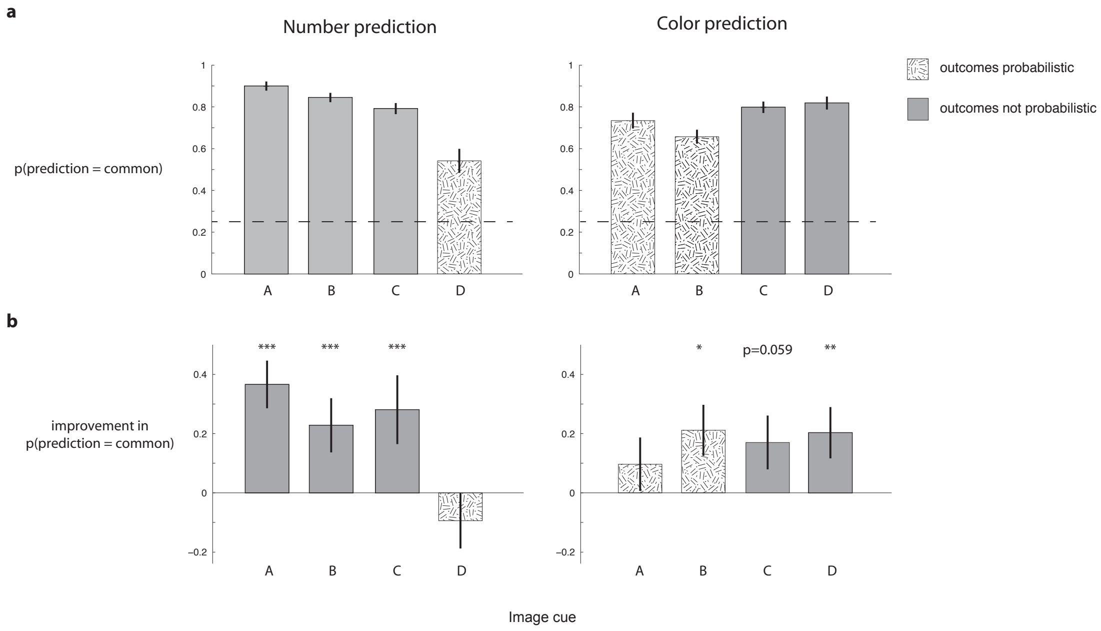


**Figure 2. Overall behavioral performance, for each image cue and prediction trial type.** Hatched bars indicate that the outcomes were probabilistic for that cue and dimension (i.e. Cue D for number, and Cues A and B for color). Error bars indicate standard error of the mean. **(a) Mean performance across the experiment.** Probability of choosing the more common outcome (the optimal prediction), for number prediction trials and color prediction trials. Dashed line indicates chance performance. **(b) Learning across the experiment.** *Change* in probability of choosing the more common outcome (computed as the difference between the last session and the training session), for number prediction trials and color prediction trials. *p < 0.05, **p < 0.01 ***p < 0.0001

In terms of learning across the experiment, the subjects tended to become more optimal in their predictions, as measured by the difference between performance on the last scan session compared to performance during the training session (before entering the scanner) (Fig 2b). The only exception was in predicting the number of M&Ms for Image D, possibly because of the high salience of the 4 M&M outcome. Improvement in color and number prediction was not significantly different for Image A and Image B.

For all prediction trial types, there was significant variance across subjects in both average performance and in learning, and so we wished to see if activity in OFC could predict that variance.

### Learning from recent outcomes during scan sessions

We evaluated each subject's sensitivity to the most recent outcome, as a behavioral proxy for learning rate – a subject with high learning rates should be relatively more likely to expect an outcome that she recently experienced, while a subject with low learning rates should be relatively unaffected by recent experience. To evaluate a subject's sensitivity to the most recent outcome, we evaluated the subjects' tendency to choose the same outcome on the next trial with the same image cue. That is, we computed the probability of the subject predicting the common outcome after most recently experiencing the common outcome, compared with after most recently experiencing the uncommon outcome. Subjects with stronger sensitivity to the

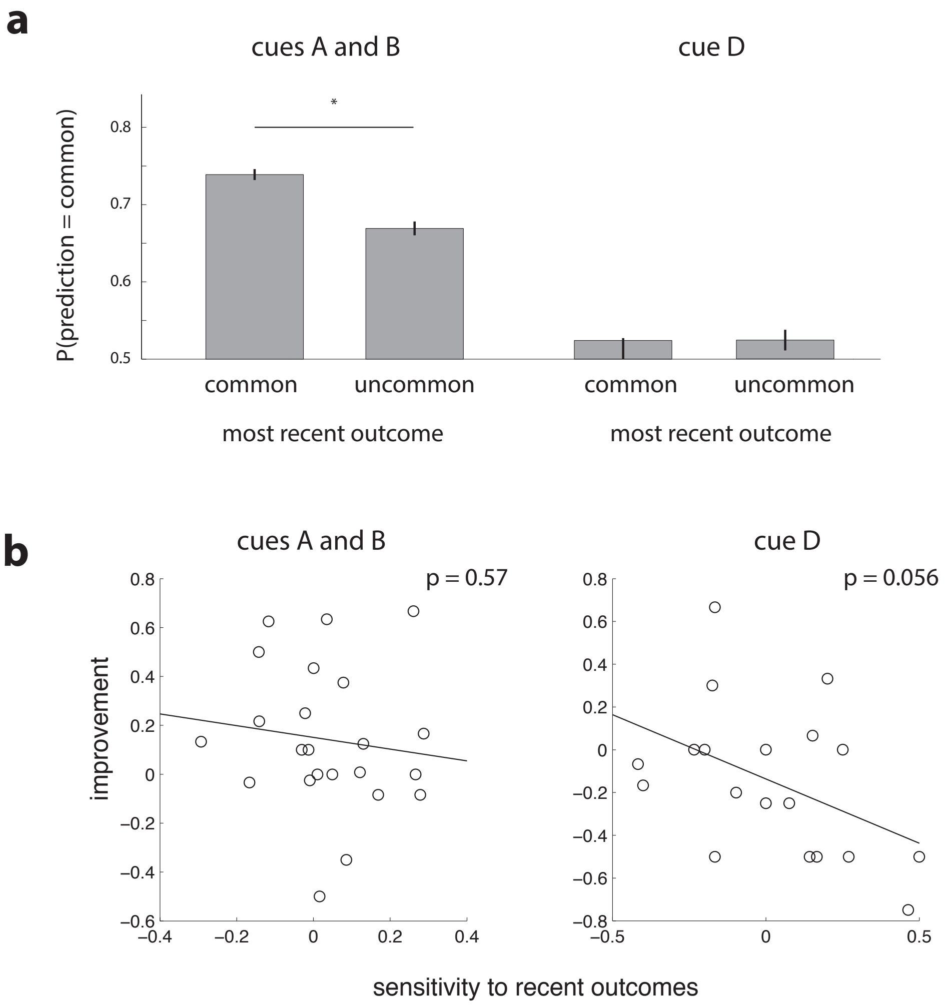


**Figure 3. Learning from recent outcomes. (a) Behavioral evidence of learning from recent outcomes during scan sessions.** Probability of predicting the common outcome, depending on whether previous outcome (for the most recent trial with the same image cue) was the common or uncommon outcome, for (left) color prediction on cue A and B trials, and (right) number prediction on cue D trials. Means ± SEM. **(b) Correlation between learning measures.** Correlations between sensitivity to recent outcomes (computed as P(prediction=common) after a common outcome – after an uncommon outcome) and improvement across the experiment (computed as the difference between P(prediction=common) for the last session and the training session).

For color prediction on Cue A and B trials (during the scan sessions), subjects showed significantly greater probability of choosing the common outcome if the most recent outcome was common, i.e. subjects were more likely to choose an outcome if they recently saw it, showing that subjects were using their experience during the scan sessions to learn about Cue A and B outcomes (Fig 3a). On average, subjects did not show this pattern of learning during the scan sessions for the Cue D number trials.

Note that higher sensitivity to recent outcomes does not imply greater improvement across the experiment, because high learning rates can in fact lead to behavior that is more random. In fact, sensitivity to recent outcomes was uncorrelated (across subjects) with improvement across the experiment for Cue A and B color prediction, and negatively correlated for Cue D number prediction (Fig 3b).

**Representation of outcomes in OFC, but not image cues**

To determine representation of the state transition itself, we used multivoxel classification methods to classify the start state (image cue) and end state (M&M outcome), at the time of the M&M outcome. Cross-validated classifier performance was significantly above chance (50%) for M&M outcome, indicating reliable representations of end state (Fig 3b). We did not find above-chance classifier performance for the start state (image cue). Note that, on each trial, the image cue was still on screen at the time that the M&M outcome appeared (and in fact occupied a much larger area of the screen than the M&Ms), indicating that representations in OFC were not a simple reflection of perceptual input.

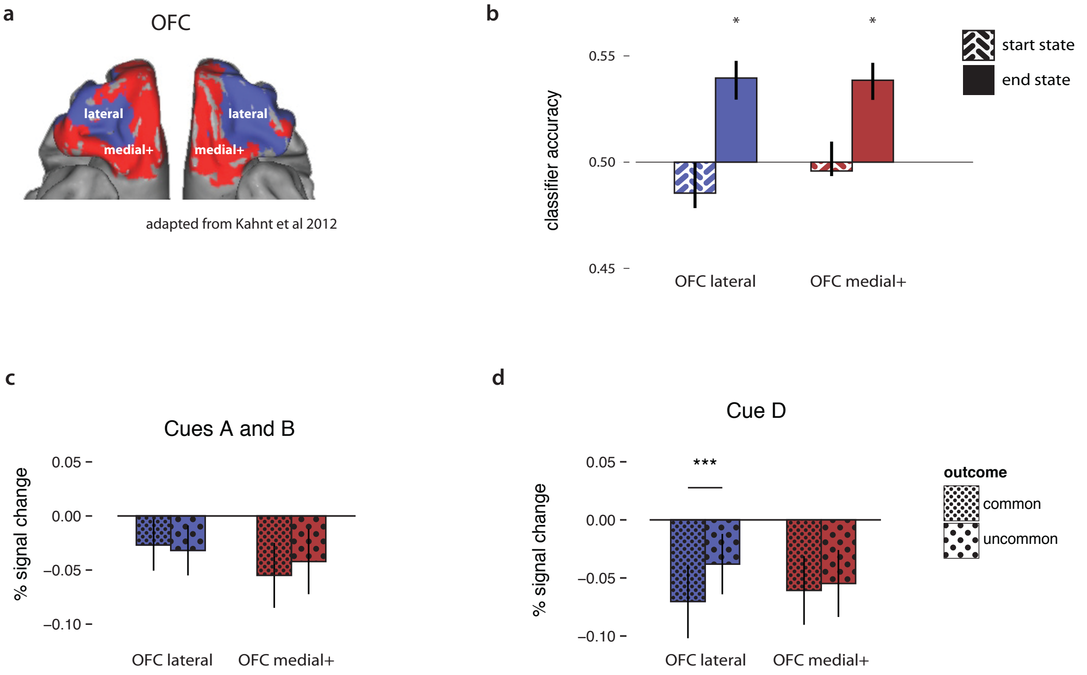


**Figure 4. Basic neural results in OFC. (a) Subregions of OFC,** displayed on the orbital surface of the brain. These regions of interest were obtained on a different dataset by Kahnt et al (2012), who parcellated the OFC using k-means clustering of functional connectivity. **(b) Classifiability of M&M outcome in OFC.** Cross-validated classification performance for start state (image cue) and end state (M&M color) for Image A and B trials, using multivariate linear classifiers on OFC activity. Mean across subjects. Error bars indicate SEM. *p < 0.05 **(c-d) Univariate responses of OFC.** Percent signal change in subregions of OFC at the time of the common outcomes and the uncommon outcomes. ***p < 0.005

## Univariate OFC responses at the time of outcome

For all cues, OFC showed negative BOLD responses at the time of the outcome ($p < 0.05$ for all cues), as has been previously observed in the OFC (e.g. Boorman et al, 2009). We did not see significant differences in univariate BOLD responses for common vs. uncommon outcomes, except in lateral OFC for Cue D (Fig. 4c-d), suggesting sensitivity to reward value or salience in lateral OFC.

### Average activity in OFC is correlated across-subjects with learning, but not performance

We found that OFC activity at the time of the outcomes for Images A and B was correlated with learning across the scan sessions for color prediction, although not in the manner we hypothesized. We had hypothesized that greater OFC responses at the time of an outcome should lead to greater change in behavior towards expecting that particular outcome. Instead, subjects with larger average negative BOLD responses in OFC during the scan sessions at the time of *any* outcome (both common and uncommon) showed a greater change in behavior towards choosing the *common* outcome for Image A and B trials, when comparing subjects' predictions before the first scan session to their predictions on the last scan session. We found this to be true for both subregions of OFC (lateral and medial+) (Fig 5a). Lateral OFC further showed a negative correlation across subjects between learning and the difference in mean activity at the time of the uncommon vs. common outcomes.

Interestingly, we did not find any relationship when comparing average activity in OFC with subjects' overall performance (Fig 5b). That is, OFC activity only showed a relationship with *change* in performance, indicating a specific role for OFC in *learning*.

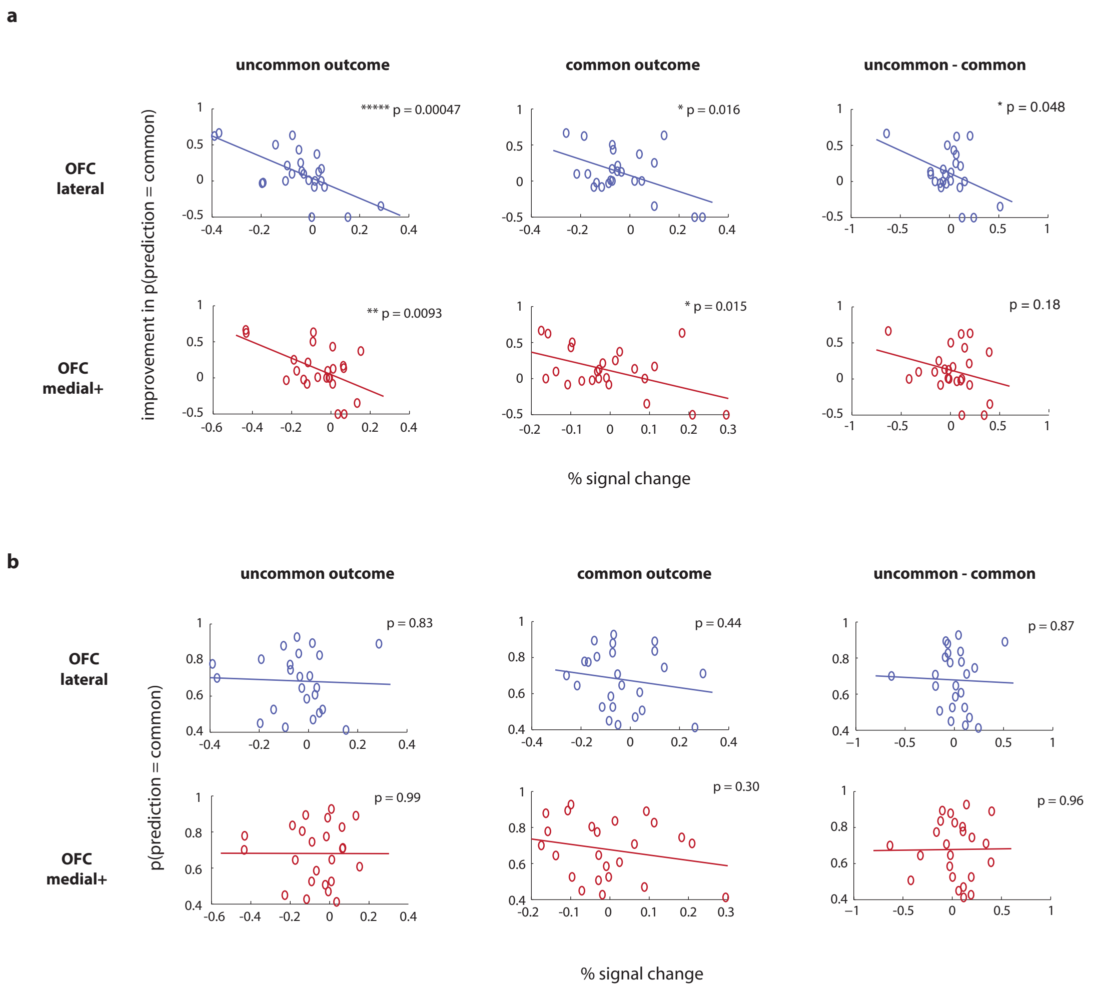


**Figure 5. Across-subject correlations of OFC activity with improvement and overall performance.** Each point indicates one subject. **(a) Across-subject correlations of OFC activity with improvement across the experiment**, measured as the change in probability of making the optimal prediction (the most common option), computed as the difference in performance between the last session and the training session. These correlations are performed for mean % signal change in OFC subregions at the uncommon outcomes and common outcomes, and also for the difference between two. **(b) Across-subject correlations of OFC activity with overall performance**, measured as the probability of making the optimal prediction (the most common option), across all sessions.

We also did not find any across-subject correlations between OFC activity and improvement for the number prediction trials (for Cues A and B, where number was held constant) or for the Cue D trials (neither number prediction or color prediction).

### Trial-by-trial correlations of OFC activity with learning from the most recent outcome

Given that subjects demonstrated learning from the most recent outcome for Cues A and B during the scan sessions (Fig 3a, described above), we evaluated whether OFC activity affected this learning, on a trial-by-trial basis. We performed logistic regression, within subjects, of OFC activity at the time of an outcome against the probability of predicting the same outcome on the next trial with the same image cue.

Again, the results indicated an involvement of OFC in learning about transitions, but not in the way we originally hypothesized. We expected that, for Cues A and B, the fitted slope terms for the logistic regression should be positive—greater OFC activity at the time of an outcome should lead to a greater probability of the subject predicting the same outcome on the next trial. Instead, we found that the fitted slope terms were positive for trials where the most recent outcome was the common outcome, and negative for trials where the most recent outcome was the uncommon outcome. In other words, no matter the outcome, greater BOLD activity in OFC at the time of an outcome was correlated with greater probability of predicting the common outcome on the next trial with the same cue (Fig 6).

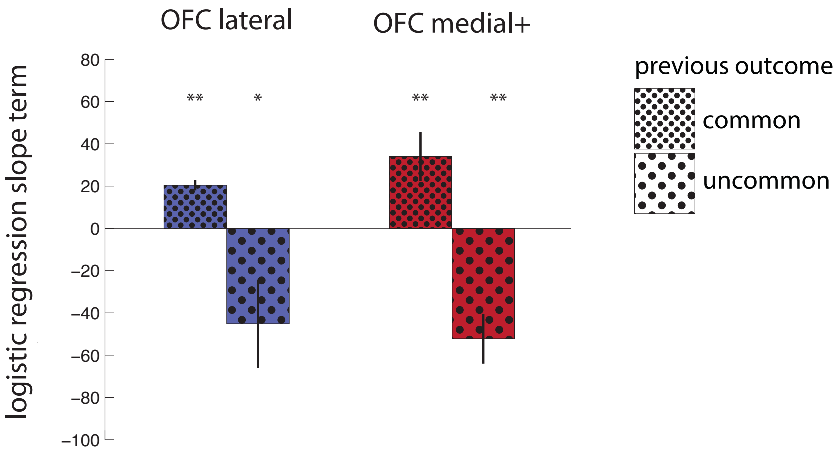


**Figure 6. Within-subject, trial-by-trial correlations of OFC activity with learning from recent outcomes, for Cue A and B trials.** Mean slope term from logistic regression of % signal change in OFC subregion at previous outcome (for the most recent trial with the same image cue) vs. probability of predicting the same outcome, fitted for each subject separately, and also separately for trials where the previous outcome was the common outcome or where the previous outcome was the uncommon outcome. Bars indicate mean slope terms across subjects ± SEM.

## 3.4 Discussion

The OFC has been shown to be particularly important for model-based decision-making, but previous work implicating OFC in the representation of expected values does not explain why it should be important for model-based decision-making in particular. Here, we have shown that OFC activity is related to learning about the transition structure of the world (the tendencies of certain states to lead to other states), an ability critical for planning, which may explain OFC importance in model-based decision-making.

Using an experimental design that facilitates constant learning of transitions between states (due to the probabilistic nature of the transitions), we have shown that activity in the OFC is correlated with behavioral measures of learning about state transitions, both within and across subjects. Across subjects, average OFC activity at the time of outcomes was negatively correlated with an improvement in optimally predicting state transitions. OFC activity was not correlated with mean performance, indicating a specific role for learning. Within subjects, on a trial-by-trial basis, OFC activity at the time of an outcome was positively correlated with a greater likelihood of optimally predicting the outcome on the next trial with the same image cue.

We can conclude that the state-transition learning in our experiment was distinct from value-based learning like that implemented in the dopaminergic system, since the number of M&Ms was always the same for the trials of interest. Also, in our analyses, we always combined conditions in which the identities (M&M colors) of the common and uncommon outcomes were reversed, so that any potential differences in value for the different colors would cancel out. Note, however, that we did find what appeared to be value sensitivity in lateral OFC for a task condition (not considered in our main analyses) in which outcomes varied in the number of M&Ms, consistent with previous work demonstrating that OFC activity encodes the value of rewards (Schoenbaum et al, 1998; Gottfried et al, 2003; Padoa-Schioppa and Assad, 2006; Hampton et al, 2006; Fellows, 2007; Hare et al, 2008; Wallis and Kennerley, 2011; Monosov and Hikosaka, 2012).

Using previously determined functional connectivity-based parcellation of OFC (Kahnt et al, 2012), we separately inspected medial and lateral OFC in all our analyses, given that previous work implicating OFC in learning about transition structure (McDannald et al, 2010) was only performed in the lateral OFC of rats (although the homology of OFC between rodents and humans is currently unclear, and also OFC subdivisions are particularly complex given observed considerable anatomical variability within individuals; Wallis et al, 2011; Chiavaras and Petrides, 2000). In our study, medial and lateral OFC showed very similar results across all our analyses. Of course, this does not rule out the possibility that there may exist a different parcellation of OFC that would lead to differing results across subregions.

Note that the BOLD signal we measured might reflect inputs to OFC, local processing within the OFC, its projections to other areas, or a combination thereof; our hypothesis does not distinguish between these interpretations. We should also take special care in interpreting the negative BOLD response to outcomes in OFC, which has been previously observed (e.g. Boorman et al, 2009) but which is not yet fully understood.

What algorithm might underlie the observed relationships between OFC and learning about state transitions? Previous work has proposed a “state prediction error” algorithm for learning state transitions, analogous to the reward prediction errors observed in dopaminergic neurons. This state prediction error should signal surprise at the time of an outcome, and would be used to adjust internal estimates of transition probabilities towards greater prediction of the observed outcome. Gläscher et al 2010 implicated some brain areas in this function (dorsolateral prefrontal cortex and intraparietal sulcus), via univariate correlation with an inferred state prediction error signal.

Our results do not uphold the idea that the OFC supports learning about state transitions via such a state prediction error signal; at the least, this signal does not seem to be encoded in the OFC's univariate response to outcomes, given that we did not observe univariate differences in OFC activity for common vs. uncommon outcomes (we would expect greater state prediction errors for the uncommon outcomes), and also the fact that OFC activity at the time of an outcome was not correlated with greater expectations of that particular outcome. Instead, OFC activity was related to greater expectations of the more common outcome, no matter whether the OFC activity occurred at a common or uncommon outcome. It is also not obvious why we should find opposite directionality in the relationship between OFC activity and learning for across-subject vs. within-subject analyses, although this result may eventually serve as a key to understanding the role of OFC. Another key to understanding may be that we observed representation in OFC of the identity of the outcomes, a result that has been previously observed (Izquierdo et al, 2004; Knutson et al, 2005; Padoa-Schioppa and Assad, 2006; Young and Shapiro, 2011; Klein-Flügge et al, 2013; Wilson et al, 2014).

In any case, however, it remains that we did observe relationships between OFC activity and learning about state transitions, both in across-subject analyses and on a trial-by-trial basis, and these observations are consistent with a previous demonstration that rats with OFC lesions were unable to learn about changes in state transitions (McDannald et al, 2011).

Although further work will be required to elucidate any learning algorithms that may underlie these relationships between OFC and learning about state transitions, the current results provides some hints for beginning to understand the involvement of OFC in this learning ability, and in model-based decision making in general.

## 3.5 References

Baxter MG, Parker A, Lindner CCC, Izquierdo AD, Murray EA. 2000. Control of Response Selection by Reinforcer Value Requires Interaction of Amygdala and Orbital Prefrontal Cortex. J Neurosci. 20:4311–4319.

Boorman ED, Behrens TEJ, Woolrich MW, Rushworth MFS. 2009. How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. Neuron. 62:733–743.

Botvinick MM, Cohen JD, Carter CS. 2004. Conflict monitoring and anterior cingulate cortex: an update. Trends in Cognitive Sciences. 8:539–546.

Carmichael S t., Price J l. 1996. Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. J Comp Neurol. 371:179–207.

Chang C-C, Lin C-J. 2011. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2:27:1–27:27.

Corbetta M, Shulman GL. 2002. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci. 3:201–215.

Daw ND, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience. 8:1704–1711.

Deichmann R, Gottfried J, Hutton C, Turner R. 2003. Optimized EPI for fMRI studies of the orbitofrontal cortex. NeuroImage. 19:430–441.

Efron B, Tibshirani R. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statist Sci. 1:54–75.

Fellows LK. 2007. The Role of Orbitofrontal Cortex in Decision Making. Annals of the New York Academy of Sciences. 1121:421–430.

Friston KJ. 2011. Functional and Effective Connectivity: A Review. Brain Connectivity. 1:13–36.

Glascher J, Daw N, Dayan P, O'Doherty JP. 2010. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron. 66:585–595.

Gottfried JA, O'Doherty J, Dolan RJ. 2003. Encoding Predictive Reward Value in Human Amygdala and Orbitofrontal Cortex. Science. 301:1104–1107.

Hare TA, O'Doherty J, Camerer CF, Schultz W, Rangel A. 2008. Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. J Neurosci. 28:5623–5630.

Izquierdo A, Suda RK, Murray EA. 2004. Bilateral Orbital Prefrontal Cortex Lesions in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward Contingency. J Neurosci. 24:7540–7548.

Kaelbling LP, Littman ML, Cassandra AR. 1998. Planning and acting in partially observable stochastic domains. Artificial intelligence. 101:99–134.

Kahnt T, Chang LJ, Park SQ, Heinzle J, Haynes J-D. 2012. Connectivity-Based Parcellation of the Human Orbitofrontal Cortex. Journal of Neuroscience. 32:6240–6250.

Klein-Flügge MC, Barron HC, Brodersen KH, Dolan RJ, Behrens TEJ. 2013. Segregated Encoding of Reward–Identity and Stimulus–Reward Associations in Human Orbitofrontal Cortex. J Neurosci. 33:3202–3211.

Knutson B, Taylor J, Kaufman M, Peterson R, Glover G. 2005. Distributed Neural Representation of Expected Value. J Neurosci. 25:4806–4812.

Kringelbach ML, Rolls ET. 2004. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Progress in Neurobiology. 72:341–372.

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. 2011. Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning. Journal of Neuroscience. 31:2700–2705.

Monosov IE, Hikosaka O. 2012. Regionally Distinct Processing of Rewards and Punishments by the Primate Ventromedial Prefrontal Cortex. J Neurosci. 32:10318–10330.

Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 16:1936–1947.

Mumford J a, Turner BO, Ashby FG, Poldrack R a. 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage. 59:2636–2643.

Öngür D, Price JL. 2000. The Organization of Networks within the Orbital and Medial Prefrontal Cortex of Rats, Monkeys and Humans. Cereb Cortex. 10:206–219.

O'Reilly JX, Woolrich MW, Behrens TEJ, Smith SM, Johansen-Berg H. 2012. Tools of the trade: psychophysiological interactions and functional connectivity. Social Cognitive and Affective Neuroscience. 7:604–609.

Padoa-Schioppa C, Assad J a. 2006. Neurons in the orbitofrontal cortex encode economic value. Nature. 441:223–226.

Polyn SM, Natu VS, Cohen JD, Norman KA. 2005. Category-Specific Cortical Activity Precedes Retrieval During Memory Search. Science. 310:1963–1966.

Rescorla RA, Wagner AR. 1972. A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In: Black AH,, Prokasy WF, editors. Classical conditioning II: current research and theory. New York: Appleton-Century-Crofts. p. 64–99.

Rudebeck PH, Murray EA. 2011. Dissociable Effects of Subtotal Lesions within the Macaque Orbital Prefrontal Cortex on Reward-Guided Behavior. Journal of Neuroscience. 31:10569–10578.

Schoenbaum G, Chiba AA, Gallagher M. 1998. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. Nat Neurosci. 1:155–159.

Selemon LD, Goldman-Rakic PS. 1988. Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: evidence for a distributed neural network subserving spatially guided behavior. The Journal of Neuroscience. 8:4049–4068.

Valentin VV, Dickinson A, O’Doherty JP. 2007. Determining the neural substrates of goal-directed learning in the human brain. J Neurosci. 27:4019–4026.

Wallis JD, Kennerley SW. 2011. Contrasting reward signals in the orbitofrontal cortex and anterior cingulate cortex. Annals of the New York Academy of Sciences. 1239:33–42.

Weissman DH, Gopalakrishnan A, Hazlett CJ, Woldorff MG. 2005. Dorsal Anterior Cingulate Cortex Resolves Conflict from Distracting Stimuli by Boosting Attention toward Relevant Events. Cereb Cortex. 15:229–237.

Wilson RC, Takahashi YK, Schoenbaum G, Niv Y. 2014. Orbitofrontal Cortex as a Cognitive Map of Task Space. Neuron. 81:267–279.

Wit S, Corlett PR, Aitken MR, Dickinson A, Fletcher PC. 2009. Differential Engagement of the Ventromedial Prefrontal Cortex by Goal-Directed and Habitual Behavior toward Food Pictures in Humans. J Neurosci. 29:11330–11338.

Young JJ, Shapiro ML. 2011. Dynamic Coding of Goal-Directed Paths by Orbital Prefrontal Cortex. J Neurosci. 31:5989–6000.

# 4 Orbitofrontal cortex represents a belief distribution over latent causes

This work was performed in collaboration with Yael Niv and Kenneth A. Norman. This work has previously presented at the following conferences: *Reinforcement Learning and Decision Making*, Edmonton, AB, Canada (2015); *Computational and Systems Neuroscience*, Salt Lake City, UT (2015); *Society for Neuroscience*, Washington, DC (2014). It is currently in submission as a journal article.

## 4.1 Introduction

In recent years, cognitive neuroscientists studying reinforcement learning have recognized the importance of specifying representations of environmental "state" that capture the structure of the world in a predictive way (Gershman and Niv, 2010; Courville et al, 2006). At the same time, there has been renewed interest among cognitive neuroscientists in how memory encoding and retrieval are shaped by situation-specific prior knowledge ("schemas", e.g. Tse et al, 2007). As work in this area progresses, it is important to clarify exactly what constitutes a schema and how schemas are formed.

Whether inferring the current "state" or the currently relevant "schema", agents are making inferences about the hidden variables that underlie and generate our observations in the world. This inference can be concretely formulated in terms of Bayesian latent cause models (e.g., Gershman, Blei, and Niv, 2010). According to this framework, states and schemas can be viewed as hidden (latent) causes that give rise to observable events. For example, if you arrive late to a lecture, the situation (whether this is indeed the department colloquium or you have accidentally walked in on an undergraduate class) affects your observations about the average age of the audience, the proportion of audience members that are taking notes, the type of information being presented, and so on. To decide whether you are in the right place, you can use Bayesian inference to infer a belief distribution over the possible situations that might have generated the current observations, i.e. a posterior probability distribution over latent causes, *p(latent cause | observations)* (Figure 1A).

We hypothesized, based on the similarity of the underlying computations, that the inference related to these two cognitive constructs (states and schemas) might be implemented using the same neural hardware. Indeed, there is one area of the brain that has separately been implicated in representing states (Wilson et al, 2014) and also schemas (Schlichting and Preston, 2015; Richards et al, 2014; Ghosh and Gilboa, 2014; Ranganath and Ritchey, 2012; van Kesteren et al, 2012; Tse et al, 2011) – the orbitofrontal cortex (OFC). Furthermore, previous univariate analyses in fMRI have implicated this region in encoding various summary statistical measures that are related to or are components of the posterior distribution, e.g. the posterior mean, likelihood of the current stimulus, and prior uncertainty (Ting et al, 2015; d'Acremont et al 2013; Vilares et al, 2012). However, these studies have not investigated representations of a full probability distribution.

Here, we used fMRI to investigate representation in OFC of posterior probability distributions over latent causes. In our experiment, we created a probabilistic environment in which participants were required to make inferences about the hidden causes that generated their observations. Participants viewed sequences of animal photographs, taken in one of four "sectors" in an animal reserve. They were tasked with judging the probability with which each sector generated the animal photographs, based on their previous experience observing animals in each sector. Using multivariate pattern similarity analyses of fMRI activity, we found that BOLD activity in the OFC was better explained by the posterior distribution over sectors (latent causes) than by a wide range of related signals, including the current stimulus, the most probable sector (the maximum a posteriori latent cause), or the uncertainty over latent causes (operationalized as the entropy of the posterior distribution). The present result advances our understanding of the function of the orbitofrontal cortex. It also unifies results from two different fields of cognitive neuroscience, inviting further investigation into the relationship between probabilistic inference, states, and schemas.

## 4.2 Methods

**Participants**

32 participants (aged 18-34 years, 22 female) from the Princeton University community participated in exchange for monetary compensation ($20 per hour + up to $15 performance-related bonus). All participants were right-handed. Participants provided informed written consent. The study was approved by the Princeton University Institutional Review Board.

**Experimental design**

***The safari***

Participants were told that they were going on a safari, visiting an animal reserve that was divided into 4 different sectors. Each sector was associated with a different color, background image, background music, and location on a 2 by 2 map (randomized across participants).

There were 5 different kinds of animals in the animal reserve. Every animal appeared in every sector, but with different likelihoods *P(animal | sector)*. The likelihoods (not shown directly to the participants) were chosen so that none of the sectors were strongly identified with a single animal, and so that none of the animals were strongly identified with a single sector (Figure 1B; colors and animals were randomly assigned across participants).

### *Procedure Overview*

The experiment consisted of two parts. In the first part, participants “toured” through the animal reserve, in order to learn (through experience) the likelihoods *P(animal | sector)* for each animal and each sector. In the second part of the experiment, participants were shown sequences of “photographs” of animals that were taken in an unknown sector, and were asked to infer the posterior probabilities of different sectors given the animals shown in each sequence, *P(sector | animals shown)*. For each participant, the experiment took place across two consecutive days (see Table 1).

| | | |
|---|---|---|
| **Day 1** | | |
| “Tours” task | 40 trials each tour | 2 tours through each sector, going clockwise around the map |
| “Tours” task | 20 trials each tour | 2 tours through each sector, sectors pseudorandomly ordered |
| **Day 2** | | |
| “Tours” task | 30 trials each tour | 2 tours through each sector, sectors pseudorandomly ordered |
| “Tours” task | 10 trials each tour | 2 tours through each sector, sectors pseudorandomly ordered |
| “Photographs” task | 2 sessions x 20 trials each | outside of the MRI scanner |
| “Photographs” task | 4 sessions x 30 trials each | inside of the MRI scanner |

***Table 1. Tasks performed by participants on Day 1 and Day 2.***

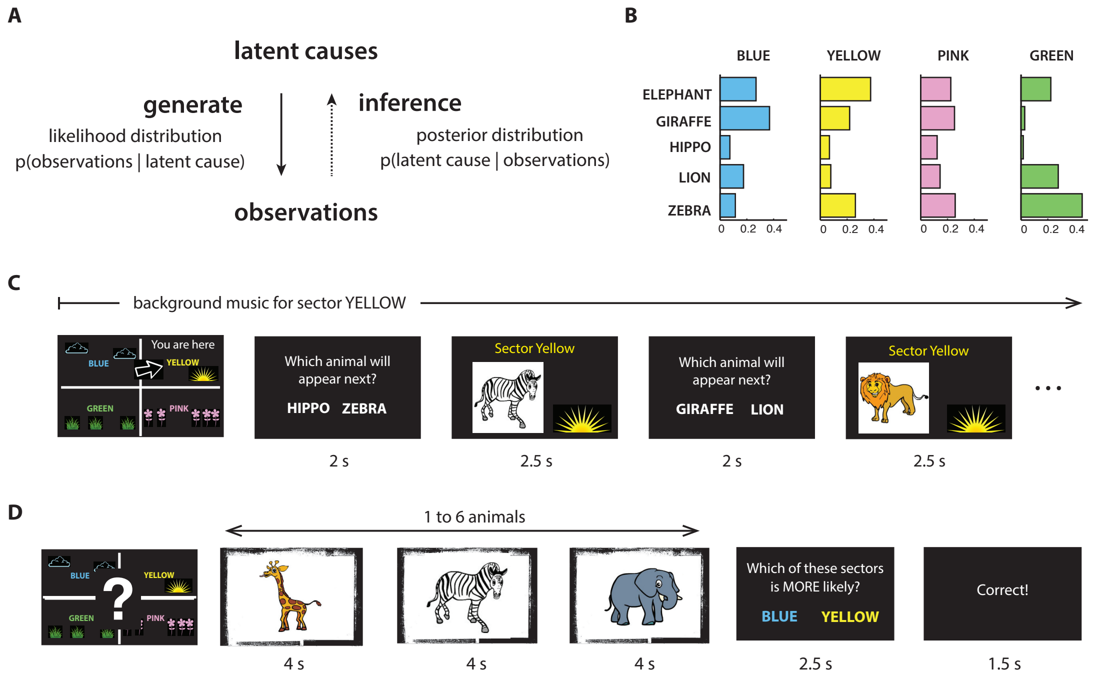


***Figure 1 – Task. A.*** *Schematic showing the relationship between latent causes and observations in the world. Inference about the posterior probability over latent causes involves inverting the generative model.* ***B.*** *Animal likelihood distributions P(animal I sector) (not shown directly to participants). Colors and animals were randomized across participants.* ***C.*** *An example of the first few trials of a tour through sector YELLOW. Each tour began with an image of the safari map, indicating the current sector and its location, and lasted 30-40 trials. Each trial began with a prompt asking the participant to guess which animal would appear next, followed by the appearance of an animal. A fixation cross was presented for 0.2-0.8 secs before each question and each animal presentation. The animals were pseudorandomly drawn from the likelihood distributions for the current sector. The sector's music played in the background, until the start of the next tour.* ***D.*** *An example of a trial in the "photographs" task. Each trial began with an image of the safari map with a question mark at its center, indicating that the current sector was unknown. Next, a sequence of 1-6 animals appeared (pseudorandomly drawn from a single sector). Finally, participants were prompted to guess which of two sectors (randomly chosen) was more (or, on half the trials, less) probable. Participants received feedback on their responses. A fixation cross was presented for the last 0.5 secs of each animal presentation. [Thanks to sciencewithme.com for the animal illustrations.]*

### *"Tours" task*

In the "tours" task (Figure 1C), participants were instructed that they would "tour" through the animal reserve, one sector at a time, in order to learn the animal frequencies in each sector (the animal likelihoods). One animal appeared on each trial, pseudorandomly chosen according to the likelihoods for that sector. Before each animal appeared, participants were shown a prompt, asking them to make a prediction about which of two animals (one correct and one randomly chosen) would appear next. The alternate (incorrect) option was chosen with uniform probability from the four other animals. To distinguish between the animals in the question prompt (which were not representative of the sector's likelihood distribution) and the animals that were actually drawn from the sector's likelihood distribution, the question prompts were shown as text while the animals drawn from the safari sector were shown as pictures.

In order for the sectors to form rich contexts, each sector was associated with a different color, background image, background music, and location on a 2 by 2 map (randomized across participants). Before the first trial of a tour through a sector, participants were shown the sector's location on the map. Also, for the duration of a tour through a sector, animals were displayed on the sector's color-matched backdrop image, and the music associated with that sector was played in the background.

### *"Photographs" task*

On each trial of the "photographs" task (Figure 1D), participants were shown a sequence of animal "photographs", without being told which sector the photographs were taken from. At the end of the sequence, participants were prompted to indicate which of two sectors (randomly chosen) was more (or less) probable. The two sector options for each question were chosen uniformly from the four sectors of the safari (and did not necessarily include the most or least likely sector). So, to perform well on the task, participants had to maintain a full posterior distribution over all four sectors (as opposed to estimating only the most probable sector, for instance).

Participants received 10 cents for every correctly answered question, and they received feedback on every trial. So that more probable sectors were not consistently associated with higher monetary value, we asked which of the two sectors was *more* probable on half of the trials, and which was *less* probable on the other half of the trials. To eliminate confounds with motor plan, the positions of the two response options were pseudorandomly assigned between left and right.

To encourage participants to update their inference of the sector probabilities after every animal presentation (as opposed to waiting until the time of the question to integrate over the animals observed), we varied the length of the sequences between 1 and 6 animals (so that the appearance of the question prompt was unpredictable), and participants were only allowed 2.5 seconds to give a response after the appearance of the question.

The posterior probability of each sector *P(sector | animals seen)* can be straightforwardly computed from the animal likelihoods, using Bayes rule (all sectors were equally likely a priori):

$$\begin{aligned} P(\text{sector } i \mid \text{animals seen}) &\propto P(\text{animals seen} \mid \text{sector } i\,) * P(\text{sector } i\,) \\ &\propto P(\text{animal 1} \mid \text{sector } i\,) * P(\text{animal 2} \mid \text{sector } i\,) * \ldots \end{aligned} \tag{Eq. 1}$$

Feedback for the responses was generated based on these posterior probabilities. Due to a bug in the code that was undetected during data collection, the feedback was incorrectly generated for some of the trials containing only one animal presentation (this affected approximately 10% of the trials). In our fMRI analyses, to account for learning from the incorrect feedback, we used each participant's estimates of the likelihoods (collected at the end of the experiment) instead of the real likelihoods, and we also performed trial-by-trial behavioral model-fitting to model learning from feedback (see next section).

Participants first performed 2 sessions (20 trials each) of the "photographs" task outside the MR scanner, to familiarize themselves with the task. They then performed 4 sessions (30 trials, approximately 11 minutes per session) inside the scanner.

### Behavioral model-fitting

To model participants' posterior inference on the "photographs" task, as well as any learning from feedback, we performed trial-by-trial model-fitting of participants' responses. We tested several classes of models:

> **Bayesian_nolearning** – This model assumed that participants were correctly computing the posterior distribution over sectors *P(sector | animals seen)* using Bayesian inference (as in Eq. 1). To obtain the model-derived likelihood of each behavioral response (and to capture stochasticity in participants' behavior), we used a softmax on the

posterior probabilities of the two options in each question prompt.

$$P(\text{response = option 1}) = \frac{1}{1+\exp[\beta * (P(\text{sector = option 1 | animals seen}) - P(\text{sector = option 2 | animals seen}))]} \quad \text{(Eq. 2)}$$

where β is an inverse temperature parameter (β = 0 implies equal likelihood for both options).

**additive** – In this model, instead of correctly multiplying the animal likelihoods together to obtain the posterior distribution over sectors (as in Eq. 1), we assumed that participants *added* the likelihoods together to obtain an “additive posterior” (normalized to sum to 1).

$$\text{"Additive posterior"} \;\propto\; P(\text{animal 1 | sector}) + P(\text{animal 2 | sector}) + \cdots \quad \text{(Eq. 3)}$$

While statistically suboptimal, we might expect this from a simple associative mechanism that brings the sectors to mind in proportion to their association strength with the animals seen. Again, to determine response probabilities, we applied a softmax operator to the additive “posterior” probabilities for the two options in each question prompt.

**most/least voter** – These models assumed that participants were only paying attention to the most common (and/or least common) animals in each sector, a similar strategy having been previously observed in a similar task (Gluck et al, 2002). During the trials, each animal appearance "voted" for (or against) the sectors in which it was the most common (or least common). To obtain the model-derived likelihood of each behavioral response, we used a softmax on the final tally at the end of each sequence.

We tested several variants of this model, e.g. tallying only the positive votes, and/or allowing an animal to “vote” for (or against) a sector if it was one of the *two* most (or least) common animals in that sector. The magnitude of the positive and negative votes were either allowed to be two separate free parameters, or constrained to be equal to each other. Because the magnitude of the vote already served as a scaling parameter for the input to the softmax operator, the inverse temperature of the softmax was kept constant at 1.

**Bayesian_feedbackRL** – These models were designed to account for learning from feedback during the “Photographs” task (including the incorrectly generated feedback). Here we assumed a reinforcement learning process, in which participants adjusted their internal estimates of the animal likelihoods after feedback about the two sectors in the question. These likelihoods were then used to compute the posterior distribution over sectors via Bayes rule.

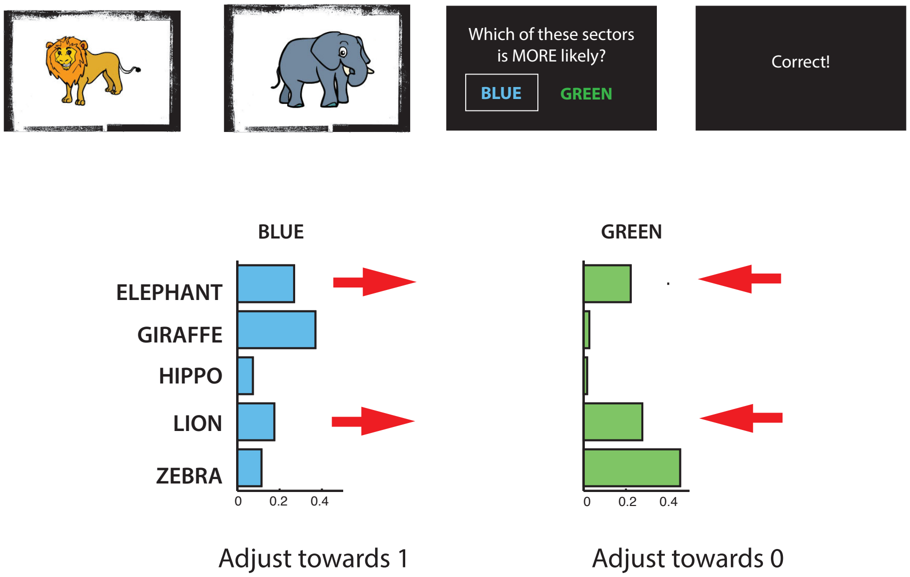


***Figure 2 – FeedbackRL model.*** *An illustration of learning from feedback in the Bayesian_feedbackRL model, for a single trial (not real data). In this example trial, the participant saw a lion and an elephant, and was asked about sector BLUE and sector GREEN. The feedback indicated that sector BLUE was more probable. As a result, the likelihoods P(BLUE | elephant) and P(BLUE | lion) are adjusted towards 1 with learning rate $\alpha_{pos}$, and the likelihoods P(GREEN | elephant) and P(GREEN | lion) are adjusted towards 0 with learning rate $\alpha_{neg}$.*

For the sector that feedback indicated to be more probable, likelihoods were adjusted upwards for all animals that were seen on that trial. For the sector that was indicated to be less probable, likelihoods were adjusted downwards for all animals seen on the trial (see Figure 2 for an example).

$$P(\text{animal} \mid \text{more probable sector})_{\text{new}} = P(\text{ animal} \mid \text{more probable sector })_{\text{old}} + \alpha_{\text{pos}}(1 - P(\text{ animal} \mid \text{more probable sector })_{\text{old}})$$

$$P(\text{animal} \mid \text{less probable sector})_{\text{new}} = (1 - \alpha_{\text{neg}})P(\text{animal} \mid \text{less probable sector})_{\text{old}}$$

(Eq. 4)

Estimates of the likelihoods were renormalized after each adjustment. The learning rates $\alpha_{pos}$ and $\alpha_{neg}$ were either allowed to be two separate free parameters, or they were constrained to be equal.

For the initialization of the likelihoods, we tested two versions of the model: initialization at the true animal likelihoods, or initialization according to the participants' subjective estimates of the likelihoods (collected at the end of the experiment, see below).

Finally, the likelihoods were used to compute the posterior distribution over sectors via Bayes rule. Thus, posterior inference in the FeedbackRL model also used Bayes rule – the only difference from the "Bayesian_nolearning" model above is that the likelihoods (which enter into the posterior inference computation from Eq. 1) were adjusted on each trial according to feedback.

We tested several additional variants of this model. In one variant, participants only adjusted their likelihoods in response to "You are incorrect" feedback (instead of in response to all feedback). In another variant of the model, we scaled the learning rates separately for each animal according to how much that animal contributed to the final posterior distribution:

$$\begin{aligned}\alpha_{\text{eff, animal X}} = \alpha \cdot abs[\, & P(\text{more probable sector} \mid \text{appearances of animal X}) \\ & - P(\text{less probable sector} \mid \text{appearances of animal X})\,] \end{aligned} \qquad \text{(Eq. 5)}$$

In this variant, animals appearing multiple times in a trial would have higher effective learning rates, having contributed more to the final decision.

In a post-experiment questionnaire, we asked participants to provide their estimates for the animal likelihoods in each sector. For each of the models above, we tested versions using (a) the *actual* animal likelihoods, and (b) *subjective estimates* of the animal likelihoods. For the few participants who provided likelihood estimates that did not sum to 1, we normalized the estimates. To avoid taking logarithms of 0, we converted estimated likelihoods of 0 to 0.01 (and renormalized).

For each of the models, we also tested versions in which the earlier and/or later animals in each sequence were given extra weight. To model these primacy/recency effects, we fit a power law function for each participant to give more weight to the earlier and/or later animals in each sequence (e.g. $1^w$, $2^w$, … for animal 1, animal 2, …). The likelihoods were exponentiated by this weighting and renormalized. If modeling both recency and primacy, the weightings for each were summed. We tested versions in which the recency and primacy free parameters *w* were either allowed to be two free parameters, or they were constrained to be equal.

Free parameters for each model were fit to each participant's behavioral data separately, using Matlab's "fmincon" function, with at least ten random initializations for each model and each participant. The best-fitting parameters (the maximum likelihood estimates) were used to evaluate, for each participant and each model, the (geometric) mean likelihood per trial (i.e., the exponentiated log likelihood per trial, without any penalization for number of parameters), the Akaike information criterion (AIC), and the Bayesian Information Criterion (BIC), in order to compare the models and determine which best accounted for participants' behavior.

| Model | Free parameters | Mean ± SE | Range |
|---|---|---|---|
| **Bayesian_nolearning** | β - softmax inverse temperature | 4.04 ± 2.22 | [0, ∞] |
| **additive** | β - softmax inverse temperature | 7.04 ± 3.46 | [0, ∞] |
| **mostleast_voter** (voting for or against the sectors in which an animal was the most or least common) | *v_pos* - size of positive vote<br>*v_neg* - size of negative vote | 1.69 ± 3.39<br>0.754 ± 1.39 | [0, ∞]<br>[0, ∞] |
| **Bayesian_feedbackRL** (learning from all feedback, no scaling of learning rates, and $\alpha_{pos} = \alpha_{neg}$) | α - learning rate<br>β - softmax inverse temperature | 0.0515 ± 0.161<br>4.82 ± 2.52 | [0, 1]<br>[0, ∞] |

***Table 2. Free parameters and parameter fits, for the best-fitting model for each class.*** *The best-fitting models for all classes did not model recency or primacy biases, and used each participant's subjective estimates of the animal likelihoods rather than the actual likelihoods. For model classes that had additional variants, the best-fitting settings are described in parentheses.*

## fMRI acquisition and pre-processing

Functional brain images were acquired using a 3T MRI scanner (Siemens, Skyra) and preprocessed using FSL (http://fsl.fmrib.ox.ac.uk/fsl/). An echoplanar imaging sequence was used to acquire 36 slices (3mm thickness with 1mm gap, repetition time (TR) = 2s, echo time (TE) = 27ms, flip angle = 71º). To increase signal in the OFC, slices were angled approximately 30 degrees from the axial plane towards a coronal orientation (Deichmann et al, 2003). For each participant, there were 4 scanning runs in total (approximately 11 minutes each). The functional images were spatially filtered using a Gaussian kernel (full width at half maximum of 5mm), and temporally filtered using a low-pass cutoff of 0.0077Hz. We performed motion correction using a six-parameter rigid body transformation to co-register functional scans, and then registered the functional scans to an anatomical scan using a 6-parameter affine transformation.

The motion regressors (and their derivatives) were residualized out from the functional images, as were the mean timecourses for cerebrospinal fluid and white matter (segmentation was performed using FSL's "FAST" function), and also the mean timecourse for blood vessels (estimated by taking voxels with the top 1% in standard deviation across time). Then, the functional images were z-scored over time. All analyses were performed for each participant in participant space, and then spatially normalized by warping each participant's anatomical image to MNI space using a 12-parameter affine transformation.

### Region of interest – Suborbital sulcus

Our region of interest (ROI) was determined as the intersection of two sets of brain areas. The first set of areas, the orbitofrontal cortex, has been postulated to be involved in the representation of “state”, due to evidence from studies of human and animal reinforcement learning (Wilson et al, 2014). The second set of areas, sometimes referred to as the “posterior medial network”, has been postulated to be involved in the computation and representation of “schemas” or “context” (Ranganath and Ritchey, 2012), as the set of areas with high connectivity with parahippocampal cortex (PHC). The intersection of these sets of areas is the suborbital sulcus, a medial subregion of the orbitofrontal cortex (Figure 6A). Using Freesurfer (Destrieux et al, 2010), the ROI was drawn as the anatomically parcellated cortical region centered on the voxel with maximal resting-state functional connectivity to PHC (Libby et al, 2012).

**Representational similarity analysis**

If the suborbital region of interest (ROI) contains a multivariate representation of the posterior distribution over latent causes, then patterns of neural activity in this area should be more similar for pairs of timepoints at which the posterior distribution was similar, and they should be dissimilar for pairs of timepoints at which the posterior distribution was dissimilar. Therefore, to test whether multivariate patterns of activity in the ROI might be representing the posterior distribution over sectors, we performed a representational similarity analysis (RSA; Kriegeskorte et al, 2008).

We first computed the similarity of the posterior distribution over sectors for every pair of timepoints during which we expected the posterior distribution to be updated (i.e. at the times of the animal appearances). This provided us with the *similarity matrix for the posterior.* We also computed the similarity of the neural pattern in the ROI for every pair of timepoints—the *similarity matrix for the ROI*. Then we computed the Spearman rank correlation of these two matrices (taking only the upper triangle and excluding the diagonal). We denote this Spearman correlation as the *similarity match between the posterior and the ROI* (Figure 3). We expected the similarity match to be positive, i.e. that the neural patterns in the ROI should be more similar for pairs of timepoints at which the posterior distribution over sectors was more similar.

We also computed the similarity match for the ROI with other signals, to compare with the similarity match between the ROI and the posterior distribution over latent causes. This is important because the similarity structure for the ROI could potentially be correlated with the similarity structure of the posterior distributions for reasons other than the fact that the posterior distribution is represented in this area. For example, the posterior distribution is, on average, more similar for pairs of timepoints at which the same animal is presented—if the suborbital ROI represents the animal currently presented, we would also find a positive similarity match between the ROI and the posterior distribution. We therefore compared the similarity match between the ROI and each alternate model, to determine the model that best explained the similarity structure of the neural data.

The set of alternate models used for this comparison included the log-transformed posterior distribution (since many signals in the brain are known to be represented in log space; e.g. Yang and Shadlen, 2010; Gibbon, 1977; Longo and Lourenco, 2007), the current stimulus, the maximum a posteriori (MAP) sector (most probable sector), the entropy of the posterior distribution

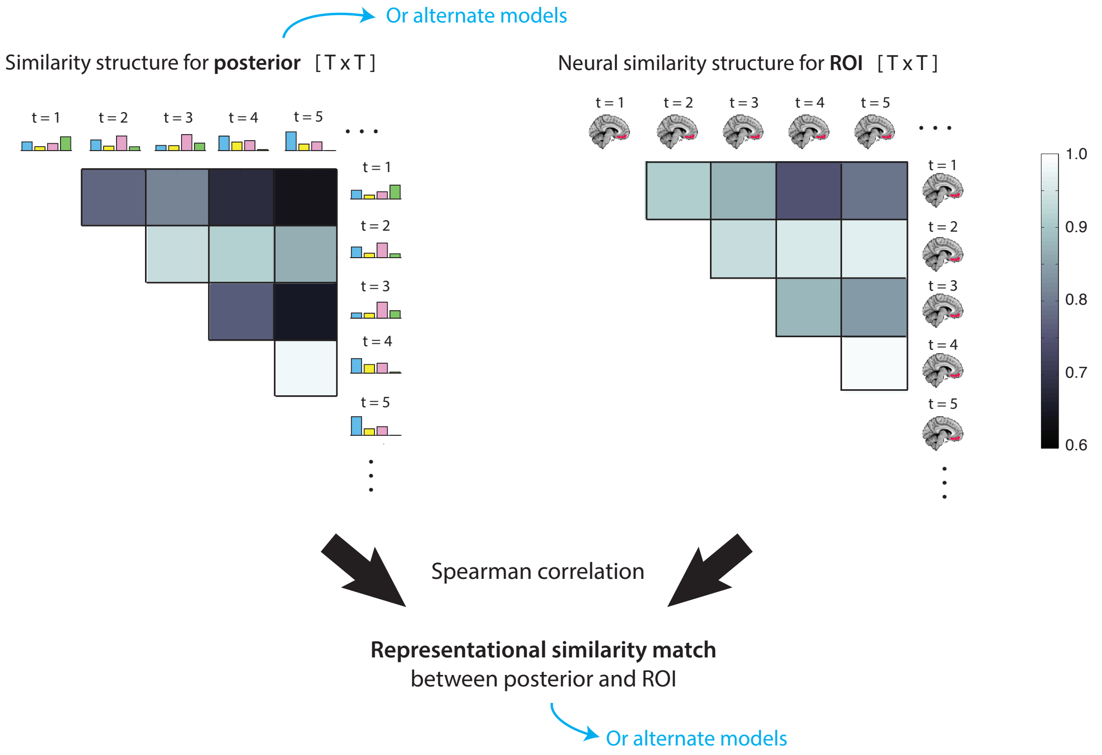

Or alternate models
Similarity structure for posterior [ T x T ]
t = 1
t = 2
t = 3
t = 4
t = 5
Neural similarity structure for ROI [ T x T ]
1.0
0.9
0.8
0.7
0.6
Spearman correlation
Representational similarity match
between posterior and ROI
Or alternate models

***Figure 3 – Representational similarity analysis.*** *An illustration of the representational similarity analysis (not real data). We first computed the similarity structure for the posterior distribution (or any alternative model; see Table 3) by computing the normalized correlation of the posterior at every timepoint with every other timepoint. We also computed the neural similarity structure for our region of interest (or for each searchlight in the whole-brain analysis), by computing the normalized correlation between patterns of activity at every timepoint with every other timepoint. To evaluate the representational similarity match between the neural data and the model, we then computed the Spearman correlation between the two matrices (using only the upper triangle of each matrix, excluding the diagonal).*

| Model | Description | Similarity measure for two timepoints |
|---|---|---|
| **posterior** | Vector [4x1] containing the posterior probability of each sector *P(sector \| animals seen so far)* | normalized correlation* |
| **log posterior** | Vector [4x1] containing the natural logarithm of the posterior probability for each sector *log[P(sector \| animals seen so far)]* | normalized correlation* |
| **current animal** | An integer ∈ $\{1,2,3,4,5\}$ indicating which animal is currently on screen | 1 if the same animal<br>0 otherwise |
| **entropy** | A scalar indicating the entropy of the posterior distribution over sectors | –abs[entropy($t_1$) – entropy($t_2$)] |
| ***maximum a posteriori*** **(MAP)** | An integer ∈ $\{1,2,3,4\}$ indicating which sector has the highest posterior probability | 1 if the same sector<br>0 otherwise |
| **p(MAP)** | A scalar indicating the probability of the maximum a posteriori sector | –abs[p(MAP($t_1$)) – p(MAP($t_2$))] |
| **posterior_MAPonly** | The posterior [4x1], zeroed for all sectors except the *maximum a posteriori* sector (i.e. a signal that contains both MAP and p(MAP) information) | normalized correlation* |
| **time** | A scalar indicating the seconds passed since the start of the session | –abs[$\text{time}_1$ – $\text{time}_2$] |
| **posterior – feedbackRL** | Vector [4x1] indicating the posterior distribution over sectors, computed using the likelihoods updated on each trial using the best-fitting feedbackRL model (free parameters fitted for each participant) | normalized correlation* |
| **MAP – feedbackRL** | An integer ∈ $\{1,2,3,4\}$ indicating the most probable sector according to the best-fitting feedbackRL model (free parameters fitted to each participant) | 1 if the same sector<br>0 otherwise |

***Table 3. Models used in the representational similarity analysis, and the similarity measure used to derive the similarity matrix.*** **The normalized correlation of vectors **x** and **y** is **x** • **y**/(||**x**|| * ||**y**||), and is equivalent to the cosine of the angle between the two vectors. It behaves differently than the more commonly used Pearson correlation; for example, the posterior distributions [0.24 0.25 0.25 0.26] and [0.26 0.25 0.25 0.24] have Pearson correlation of -1 but normalized correlation of 0.9994. We used normalized correlation because this measure accords better with intuition regarding the similarity of posterior distributions and quantities derived from posterior distributions; however, similar results were observed when using Pearson correlations instead.*

(a proxy for overall uncertainty), the probability of the maximum a posteriori sector (a proxy for confidence, acting approximately as the converse of the entropy), and temporal distance between measurements (because fMRI pattern similarity is known to vary as a function of the temporal distance between measurements). We also included models of the posterior and MAP that were instead derived using the Bayesian_feedbackRL model (given that this was the best inference model after Bayesian_nolearning, as determined from behavioral model-fitting, described above). See Table 3 for a full list of models tested.

To investigate the specificity of the result to our region of interest, we also performed a whole-brain "searchlight" analysis, using 25-voxel spherical searchlights. As with the region of interest, we computed the similarity of the neural patterns in each searchlight, to obtain the *neural similarity matrix for the searchlight.* We then computed the Spearman correlation of the similarity matrix for each searchlight with each of our models. The analysis was repeated for a searchlight centered on every voxel in the brain.

For both the ROI and searchlight analyses, the neural pattern for each animal appearance was averaged over the two TRs during which the animal appeared on the screen (after correcting for the hemodynamic lag with a 4 second shift). Similarity for neural patterns was computed using normalized correlation, to accord with the similarity measure used for the posterior-based models (similar results are obtained when using Pearson correlation instead).

Searchlight results are displayed on an inflated brain, using the AFNI SUMA surface mapper (http://afni.nimh.nih.gov/afni/suma).

### *Statistics and confidence intervals*

Unless stated otherwise, all statistics were computed using random-effects bootstrap distributions on the mean by resampling participants with replacement (Efron & Tibshirani, 1986). All confidence intervals in the text are given as standard error of the mean.

To test the reliability of searchlight results across participants, we used the “randomise” function in FSL (http://fsl.fmrib.ox.ac. uk/fsl/fslwiki/randomise) to perform permutation tests and generate a null distribution of cluster masses for multiple comparisons correction (using FSL’s “threshold-free cluster enhancement”, $P < 0.05$ two tailed).

## 4.3 Results

### **Participants learned the animal likelihoods in the “Tours”**

We evaluated participants’ final learning of the likelihood of each animal in each sector using performance from the last set of tours on the last day. In those tours, the participants chose the more likely animal 73 ± 3% of the time. Note that even if participants had perfect knowledge of the animal likelihoods, we would not expect participants to choose the more likely animal 100% of the time, due to probability matching, the well-documented behavior in which humans and animals match their choice probabilities to the probability of each option being correct, rather than choosing the most likely option every time (e.g. Vulkan, 2000; Erev and Barron, 2000). With perfect knowledge of the animal likelihoods and a probability matching policy, participants would be expected to choose the more likely animal only 69% of the time.

In a post-experiment questionnaire, we asked participants to estimate the animal likelihoods *P(animal* | *sector)* for every animal and every sector. These estimates were close to the true likelihoods, on average (Figure 4A). The mean KL-divergence of the estimated likelihoods from the real likelihoods was 0.13 ± 0.015. As discussed below, we used these participant-estimated likelihoods in our neural analyses, in lieu of the correct likelihoods.

### Performance on “Photographs” task suggested maintenance of posterior distributions over sectors

During the fMRI scan sessions, participants correctly chose the more (or less) probable sector 67 ± 1% of the time, which is significantly above chance (t-test $p$ < 1e-12). Moreover, logistic regression on participants’ responses showed that, the larger the difference in posterior probability between the correct and incorrect options, the more likely participants were to choose the

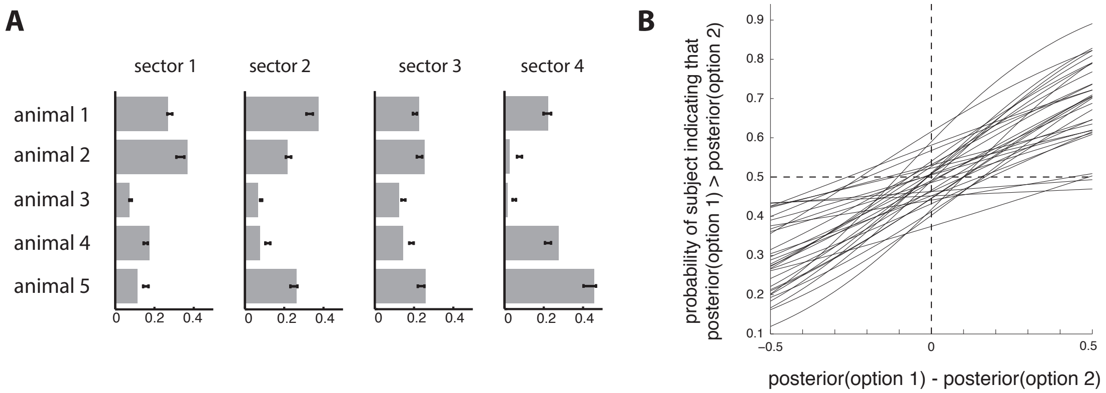


***Figure 4 – Behavioral performance. A.** Participants' subjective estimates of the animal likelihoods P(animal I sector), for each animal and each sector, collected in a post-experiment questionnaire. Gray bars indicate the true likelihoods, black intervals indicate the mean estimates ± SEM. **B.** Logistic regression on participants' responses during the fMRI scan sessions suggests that participants learned and utilized the full posterior distributions (each line shows logistic regression for one participant). The x-axis indicates the difference in posterior probability between the first and second options in the question. The y-axis indicates the probability that a participant would indicate that the first option has higher posterior probability than the second option. Mean regression parameters across participants: slope = 1.8 ± 0.040, intercept = -0.04 ± 0.15.*

correct answer (Figure 4B). Again, as in the Tours task, we expected stochasticity in participants' behavior due to probability matching. With perfect probability matching and perfect inference of the sector posteriors, we would expect participants to choose the correct option 73% of the time.

Note that the two sector options in each question were chosen at random, and therefore required participants to discriminate between posterior probabilities for any possible pair of sectors. Interestingly, participants performed similarly well whether or not questions included the maximum a posteriori (MAP; most probable) sector (accuracy 69 ± 2% for questions including the MAP, 66 ± 1% for questions not including the MAP; not significantly different). This result further indicates that participants were tracking the full posterior distribution, and not just the MAP sector.

### Trial-by-trial behavioral model-fitting suggested that participants were approximately Bayesian

The relative performance of the behavioral models is shown in Figure 5, and the mean parameter fits are shown in Table 2. For model comparison, we used the best-performing version from each class of models (these settings described in Table 2).

The two Bayesian models (with and without feedbackRL) performed best, explaining the data about equally well. Overall, the model with feedbackRL was the best model according to AIC, but the Bayesian model without learning was the best model according to BIC, which penalizes more strongly for extra parameters.

The additive model performed worse than the Bayesian models, indicating that participants were accumulating evidence multiplicatively, in accordance with the optimal strategy (Eq. 1). None of the heuristic inference models that we tested (the most-least voter class of models) could successfully outperform the Bayesian models. Nor did we identify any significant effect of recency or primacy (any small improvements in the model likelihoods were not justified by the increased number of parameters). We therefore concluded that participants were Bayesian or near-Bayesian in their inference.

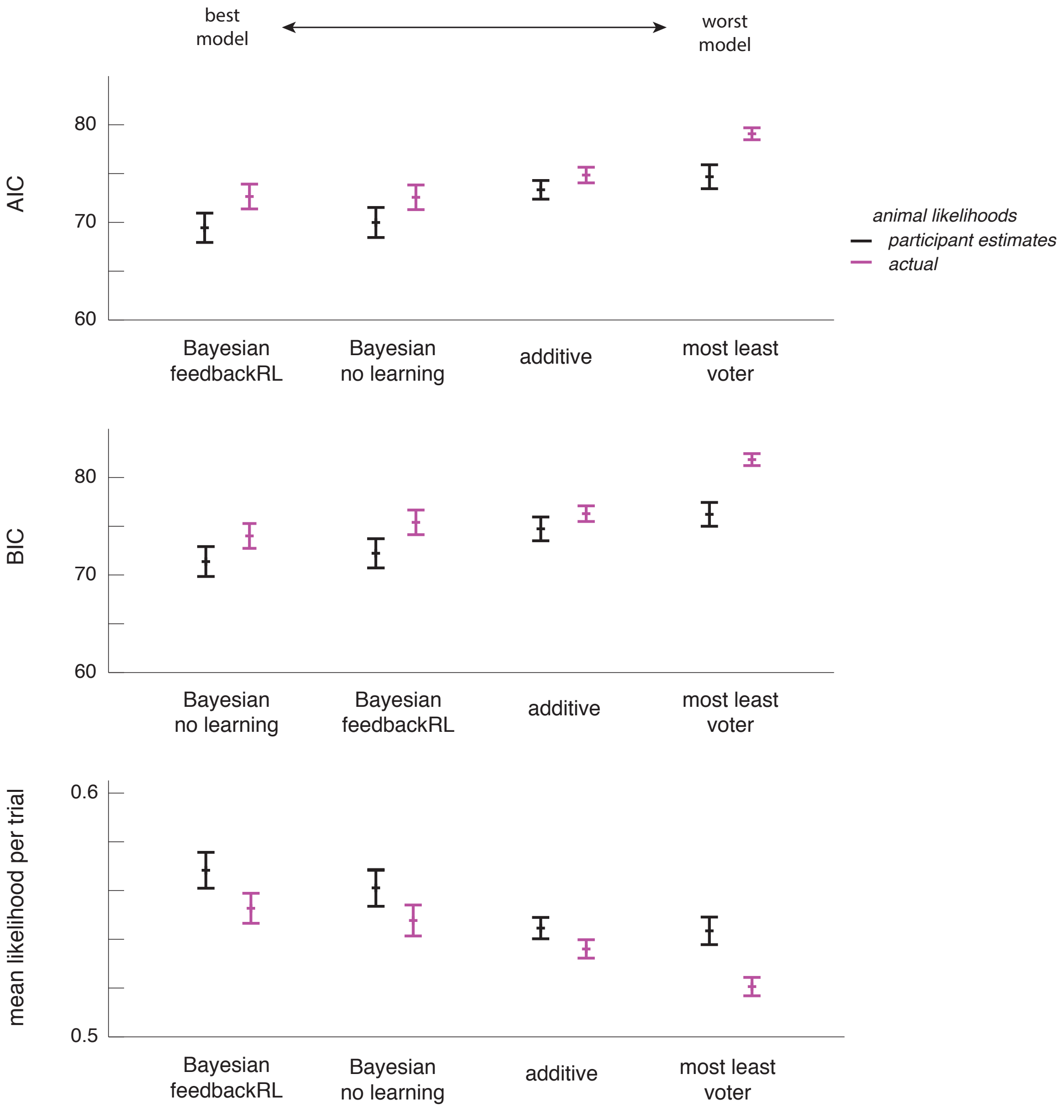


***Figure 5 – Behavioral model-fitting.*** *Akaike information criterion (AIC), Bayes information criterion (BIC), and (geometric) mean likelihood per trial (i.e. the exponentiated mean log likelihood per trial, without penalization for number of parameters) for the best-fitting model in each class (mean ± SEM across participants) suggest that the Bayesian models explained the behavioral data best. Note that better model fits are indicated by low AIC and BIC scores, but high mean likelihood. Results are shown for model fits using the participant estimates of the likelihoods or using the actual (true) likelihoods.*

As shown in Figure 5, using the participants' subjective estimates of the animal likelihoods (from the post-experiment questionnaire) provided a better fit for all models, as compared to using the real animal likelihoods. This may be surprising for the feedbackRL model, given that the participant estimates were elicited at the end of the experiment, but were used in the model to initialize estimates of the likelihoods. However, the low learning rates (see Table 2 for average fit learning rates; also, 19% of participants had fitted learning rates of 0) suggest that changes in the likelihoods throughout the experiment were small relative to the differences between the real and estimated likelihoods. The low learning rates also explain why the feedbackRL model fit the data similarly well to a Bayesian model that did not allow for changes of the likelihood during the task – the models are nested (identical for learning rates of zero) and similar for low learning rates.

### Representational similarity analysis suggests that suborbital sulcus contains a representation of the (log) posterior distribution over latent causes

Figure 6B shows the representational similarity match of the suborbital sulcus with each of the models, relative to the representational similarity match with the best model – the logposterior. For all of the alternative models tested, 95% or more of our bootstrap samples showed better representational similarity match for the logposterior than for the alternative model.

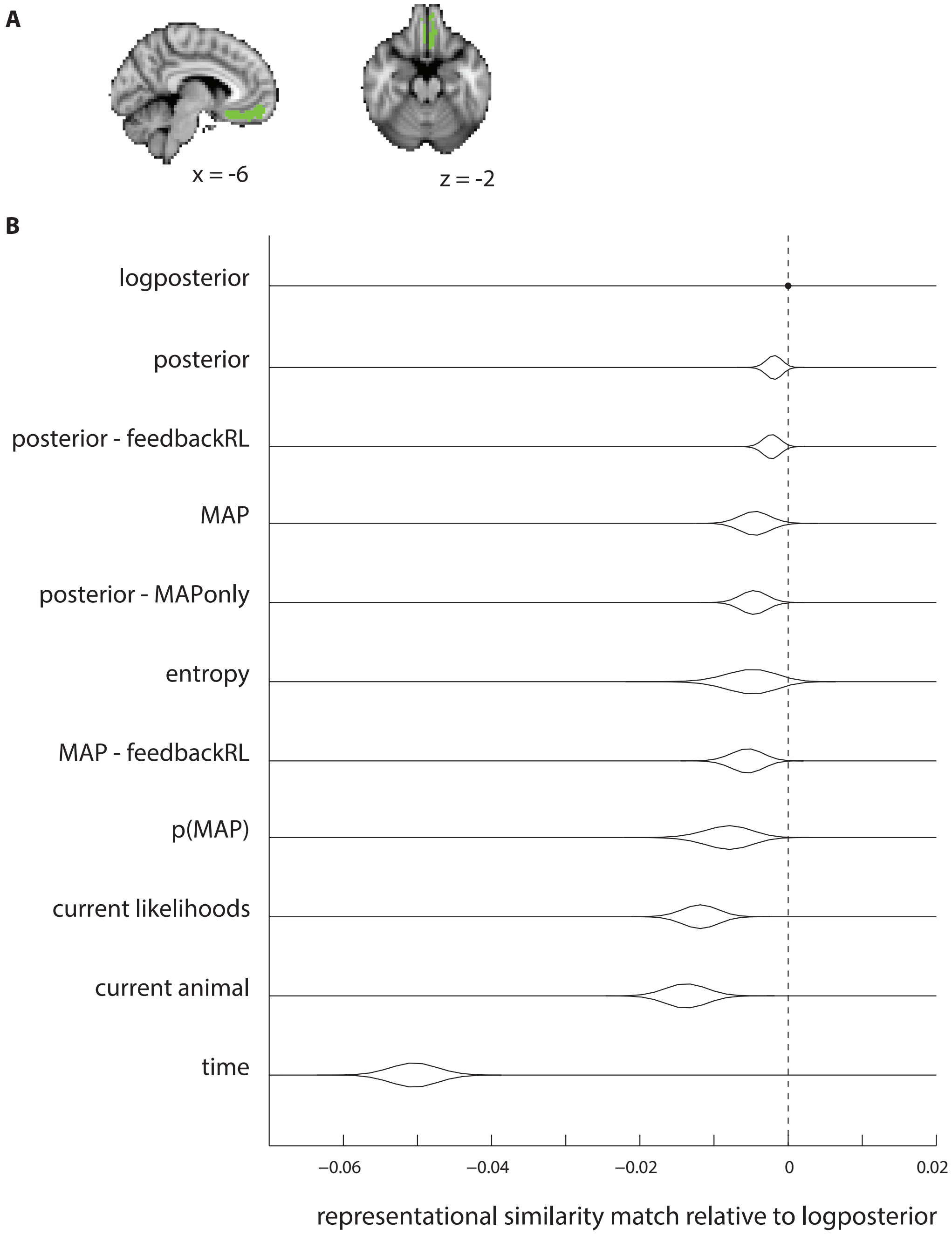


***Figure 6 – Representational similarity match for each model in the ROI. A.*** *Region of interest – the suborbital cortex, a medial subregion of the orbitofrontal cortex (OFC). See Methods for a description of how the region was defined.* ***B.*** *Representational similarity match for each of the models tested (Table 3), relative to the best model for the data (the logposterior), ordered by mean representational similarity match. The logposterior model showed the highest mean representational similarity match. The plots show bootstrap distributions on the within-participant differences, for each of the models compared with the logposterior. For all of the alternative models tested, 95% or more of our bootstrap samples showed a better match for the logposterior than for the alternative model.*

Because the posterior distribution tends to be more similar for neighboring timepoints compared with more distant timepoints, and that might also be the case for neural patterns, we took special care to verify that the logposterior model was superior to the alternative (control) time model. This was indeed the case. Moreover, we found that the temporal model displayed negative representational similarity match with the neural patterns, because BOLD patterns for neighboring timepoints tended to be anti-correlated. This result was not dependent on our linear model for temporal distances—because we used Spearman's rank correlation to compute representational similarity match, the negative similarity match result would be observed for any other model of temporal distance that falls off monotonically (e.g. an exponential model). Therefore, since the posterior distribution showed positive similarity match while the temporal model showed negative similarity match, we can conclude that any positive correlations between the similarity matrices for the posterior distribution and time cannot be responsible for the representational similarity result for the posterior distribution.

Searchlight results for the representational similarity analysis are shown in Figure 7. The orbitofrontal and ventromedial prefrontal cortex showed significantly greater representational similarity match for the logposterior model compared to every other model ($p < 0.05$ corrected, for every comparison), except for the entropy and the posterior models. It also showed greater representational similarity match for the logposterior than entropy using a more liberal threshold of $p < 0.05$ uncorrected.

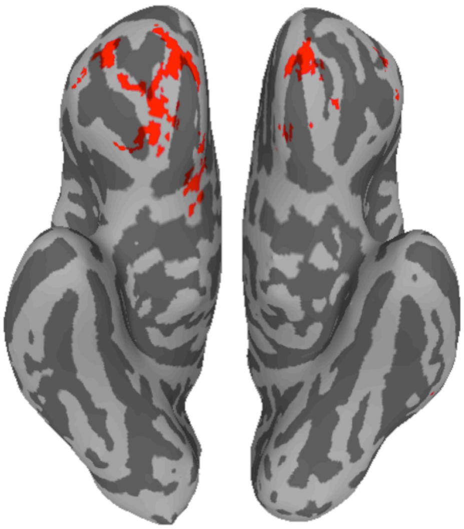


***Figure 7 – Wholebrain searchlight result.*** *Brain areas that passed both of the following criteria: (1) significantly higher representational similarity match with the logposterior model as compared with every other model from Table 3 except the posterior, the posterior from the feedbackRL model, and the entropy, at $p < 0.05$ with whole-brain correction for every comparison; (2) higher representational similarity match with the logposterior compared to the entropy, at $p < 0.05$ uncorrected. The map is displayed on the orbital/ventral surface of an inflated brain.*

## 4.4 Discussion

Because the underlying structure of the world is often not directly observable, we must make inferences about the underlying situations or “latent causes” that generate our observations. The statistically optimal way to do this is to use Bayes rule to infer the posterior distribution over latent causes. Based on previous studies implicating the orbitofrontal cortex (OFC) in the representation of the current context or situation (related to the ideas of “state” in studies of reinforcement learning, and “schemas” in studies of episodic memory), we hypothesized that the OFC might represent a posterior probability distribution over latent causes, computed using approximately Bayesian inference. To test this, we asked participants to make inferences about the probability of possible situations, in an environment where the situation probabilistically generated their observations.

Using representational similarity analysis of fMRI activity during the inference task, we found that patterns of activity in the suborbital sulcus within the OFC were indeed best explained as representing a posterior distribution over latent causes. Searchlight analyses implicated OFC more generally in this representation. Furthermore, participants' behavioral performance showed that they had access to a full posterior distribution over the latent causes for their choices; using trial-by-trial model fitting, we showed that participants' behavior was best explained as using Bayesian inference.

Our study provides evidence that the OFC represents a full posterior distribution over situations, as opposed to the best guess of the situation (the maximum a posteriori; MAP) or other summary measures of the distribution such as the overall uncertainty. We operationalized uncertainty as the entropy of the distribution—the highest entropy occurs when the distribution is completely flat (i.e., the participant is maximally uncertain about which latent cause generated the observations), and the lowest entropy occurs when the distribution is fully loaded on one latent cause (i.e., the participant is absolutely certain about which latent cause generated the observations). Our similarity analyses showed the entropy to have widespread positive similarity match in many areas of cortex, which we might expect because entropy should be correlated with the difficulty of the task, and so entropy might therefore be correlated with greater overall activity in many regions of the brain. Nonetheless, in greater than 95% of our bootstrap samples, activity in the OFC was better explained by the posterior distribution than by the entropy. Furthermore, searchlight analyses showed the specificity of this result.

Our results, using multivariate analysis, build on previous fMRI studies that have used univariate analyses in OFC to investigate a range of summary statistical quantities that are related to the posterior distribution, but which do not capture the full distribution. These studies have shown that univariate activation of the ventromedial PFC (which includes or is similar to our ROI) is correlated with a variety of summary statistics, e.g. expected reward (Ting et al, 2015), reward uncertainty (Tobler et al, 2007; Critchley et al, 2001), variance of the prior distribution in a sensory task (Vilares et al, 2012), and marginal likelihood of the current stimulus (d'Acremont et al 2013). Our experiment employed several key features — (a) multivariate neural analysis (b) four different latent causes, and (c) dissociation of latent cause from both reward and motor plan — that allowed us to identify orbitofrontal representation of a full posterior distribution over latent causes that was separate from value, and which explained neural activity in the area better than any single summary statistic that we tried. Our result may therefore explain why evidence for different summary statistics was found in different studies—these are all components of the full posterior distribution, or correlates of it.

Our study also builds on previous work in the fields of reinforcement learning and episodic memory that has implicated the OFC in representations of the current situation or context. In reinforcement learning, a belief distribution over states is necessary for optimal decision-making when the state of the world is not directly observable (partially observable Markov decision processes; Kaelbling et al, 1998). The OFC has long been implicated in reinforcement learning and decision-making in a wide range of settings; a recent review provides a unifying explanation for these results by postulating that the OFC represents inferred states in partially observable situations (Wilson et al, 2014). In theories of episodic memory, it is believed that we organize our memories according to an inferred “schema” that specifies the situation and stores previously learned relationships that a new memory can be incorporated into (Tse et al, 2007; Hupbach et al, 2008). These schemas seem to be represented or processed in the ventromedial prefrontal cortex (vmPFC, an area of the brain that is similar to our ROI; for reviews, see Schlichting and Preston, 2015; van Kesteren et al, 2012; Ranganath and Ritchey, 2012). For example, Tse et al (2011) showed evidence that activation of rat mPFC is highest immediately after memory encoding that should involve incorporating new information into existing schemas. Ezzyat and Davachi (2011) showed that greater activation of ventromedial PFC in humans during memory encoding is correlated with how strongly those memories are associated with other memories in the same “event”, consistent with the idea that vmPFC is involved in schemas that are bound to memories.

Our results confirm the involvement of OFC in representations of the current situation, and additionally show that this representation in OFC takes the form of a *distribution* over possible situations.

Finally, our work also builds on previous studies investigating neural circuits involved in the “weather prediction” task, very similar to ours, in which one of two “weather” outcomes is probabilistically predicted by sequences of cards. Knowlton et al (1996) implicated the striatum in the learning of these probabilistic associations. In our task, participants learned the animal likelihoods outside the MR scanner, and thus we could not assess the brain areas involved in the learning phase. However, our results are compatible with Knowlton et al’s insofar as the OFC may use associations learned by the striatum (in our experiment, the animal likelihoods) to make inferences when presented with new observations (in our experiment, the “photographs” task). More recently, Yang and Shadlen (2007) used the weather-prediction task to show representation of a decision variable in parietal cortex that took the form of the log likelihood ratio between two options. In our experiment, we decorrelated the posterior probability from both decision variables and stimulus-reward associations, and we also investigated representations of the posterior probability over latent causes *before* the decision period. We conjecture that the OFC contains representations of the current state or situation in terms of a posterior distribution over the possible states, a representation that is likely used by downstream areas, e.g. parietal cortex, for decision making.

Previous work on the weather-prediction task also showed that most individuals employed heuristic strategies in inferring the weather (Gluck et al, 2002). In our experiment, we explored several heuristic models of participants' inference, but were not able to find any that predicted participants' behavior better than the optimal Bayesian models. There are several reasons why our task may have discouraged the use of heuristics. First, the animal likelihoods in our experiment were designed to avoid one-to-one mappings between observations and latent causes. Second, the task environment had four possible latent causes (instead of two), and the task itself required rank-ordering all four latent causes rather than just estimating the maximum a posteriori, thus increasing complexity and leading to the inadequacy of simple heuristics. Finally, we provided participants with a large amount of training on the probabilistic model of the world, so that heuristics may have been less necessary.

The posterior distribution we found in the OFC was best modeled as being represented in log space. Representation in log space may be advantageous because addition can then replace the multiplicative operation required to accumulate evidence in non-log space (e.g. across animal presentations, in our experiment); the ability of neurons to add is well-characterized, while it is less clear to what extent neurons can multiply (Yuste and Tank, 1996; Peña and Konishi, 2001; Gabbiani et al, 2002). Indeed, neural representation in log space is common in many domains, e.g. decision variables (Yang and Shadlen, 2010), time (Gibbon, 1977) and numbers (Longo and Lourenco, 2007).

To summarize, we designed a task in which participants' observations were probabilistically generated by unobserved "situations" or "latent causes", and found evidence that OFC represents a probability distribution over possible latent causes. A representation of the log posterior distribution explained OFC activity better than alternatives such as the best guess of the current situation, or overall uncertainty in the current situation. This finding was further supported by behavioral evidence that participants had access to the full probability distribution for decision-making, and used Bayesian inference to compute the probability distribution. Our results may explain why previous studies of OFC have found evidence for representation of various summary statistical quantities in OFC (these are in fact components of the full posterior probability distribution). Our results may also unify findings from disparate literatures on reinforcement learning and episodic memory, which separately implicate the OFC in representations of the current situation.

## 4.5 References

Courville AC, Gordon GJ, Touretzky DS, Daw, ND (2003) Model uncertainty in classical conditioning, in: Advances in Neural Information Processing Systems 16:977-984.

Critchley HD, Mathias CJ, Dolan RJ (2001) Neural Activity in the Human Brain Relating to Uncertainty and Arousal during Anticipation. Neuron 29:537–545.

d'Acremont M, Fornari E, Bossaerts P (2013) Activity in Inferior Parietal and Medial Prefrontal Cortex Signals the Accumulation of Evidence in a Probability Learning Task. PLoS Comput Biol 9, e1002895.

Deichmann R, Gottfried J, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. NeuroImage 19:430–441.

Destrieux, C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. NeuroImage 53:1–15.

Efron B, Tibshirani R (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statist. Sci. 1:54–75.

Erev I, Barron G (2005) On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. Psychological Review 112:912–931.

Ezzyat Y, Davachi L (2011) What Constitutes an Episode in Episodic Memory? Psychological Science 22:243–252.

Gabbiani F, Krapp HG, Koch C, Laurent G (2002) Multiplicative computation in a visual neuron sensitive to looming. Nature 420:320–324.

Gershman SJ, Blei DM, Niv Y (2010) Context, learning, and extinction. Psychological review 117:197–209.

Gershman SJ, Niv Y (2010) Learning latent structure: carving nature at its joints. Current opinion in neurobiology 20:251–6.

Ghosh VE, Gilboa A (2014) What is a memory schema? A historical perspective on current neuroscience literature. Neuropsychologia 53:104–114.

Gibbon J (1977) Scalar expectancy theory and Weber's law in animal timing. Psychological Review 84:279–325.

Gluck MA, Shohamy D, Myers C (2002) How do People Solve the "Weather Prediction" Task?: Individual Variability in Strategies for Probabilistic Category Learning. Learn. Mem. 9:408–418.

Gold JI, Shadlen MN (2002) Banburismus and the Brain: Decoding the Relationship between Sensory Stimuli, Decisions, and Reward. Neuron 36:299–308.

Hupbach A, Hardt O, Gomez R, Nadel L (2008) The dynamics of memory: Context-dependent updating. Learn. Mem. 15:574–579.

Kaelbling LP, Littman ML, Cassandra AR, 1998. Planning and acting in partially observable stochastic domains. Artificial intelligence 101:99–134.

Knowlton BJ, Mangels JA, Squire LR (1996) A Neostriatal Habit Learning System in Humans. Science 273:1399–1402.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in systems neuroscience 2:4.

Libby LA, Ekstrom AD, Ragland JD, Ranganath C (2012) Differential Connectivity of Perirhinal and Parahippocampal Cortices within Human Hippocampal Subregions Revealed by High-Resolution Functional Imaging. Journal of Neuroscience 32:6550–6560.

Ranganath C, Ritchey M (2012) Two cortical systems for memory-guided behaviour. Nature Reviews Neuroscience 13:713–726.

Richards BA, Xia F, Santoro A, Husse J, Woodin MA, Josselyn SA, Frankland PW (2014) Patterns across multiple memories are identified over time. Nat Neurosci 17:981–986.

Schlichting ML, Preston AR (2015) Memory integration: neural mechanisms and implications for behavior. Current Opinion in Behavioral Sciences 1:1–8.

Ting C, Yu C, Maloney LT, Wu S (2015) Neural Mechanisms for Integrating Prior Knowledge and Likelihood in Value-Based Probabilistic Inference. J. Neurosci. 35:1792–1805.

Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2007) Reward Value Coding Distinct From Risk Attitude-Related Uncertainty Coding in Human Reward Systems. Journal of Neurophysiology 97:1621–1632.

Tse D, Langston RF, Kakeyama M, Bethus I, Spooner PA, Wood ER, Witter MP, Morris RGM (2007) Schemas and Memory Consolidation. Science 316:76–82.

Tse D, Takeuchi T, Kakeyama M, Kajii Y, Okuno H, Tohyama C, Bito H, Morris RGM (2011) Schema-Dependent Gene Activation and Memory Encoding in Neocortex. Science 333:891–895.

van Kesteren MTR, Ruiter DJ, Fernández G, Henson RN (2012) How schema and novelty augment memory formation. Trends in Neurosciences 35:211–219.

Vilares I, Howard JD, Fernandes HL, Gottfried JA, Kording KP (2012) Differential Representations of Prior and Likelihood Uncertainty in the Human Brain. Current Biology 22:1641–1648.

Vulkan N (2000) An economist's perspective on probability matching. Journal of economic surveys 14:101–118.

Wilson RC, Takahashi YK, Schoenbaum G, Niv Y (2014) Orbitofrontal Cortex as a Cognitive Map of Task Space. Neuron 81:267–279.

Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. Nature 447:1075–1080.

Yuste R, Tank DW (1996) Dendritic integration in mammalian neurons a century after Cajal. Neuron 16:701–716.

# 5 Slow time scales of meaning influence recall organization

This work was performed in collaboration with Marissa A. Applegate and Kenneth A. Norman. This work has previously presented at the following conferences: *Context and Episodic Memory Symposium*, Philadelphia, PA (2014); *Society for Neuroscience*, San Diego, CA (2013); *Context and Episodic Memory Symposium*, Philadelphia, PA (2013); *Cognitive Neurosceince Society Annual Meeting*, San Francisco, CA (2013). It is currently in preparation as a journal article.

## 5.1 Introduction

We have an immense number of memories stored in our brains. Why do we retrieve certain memories at certain times? How are memories organized in the brain and how does this affect recall? These questions have been studied using memory tests such as free recall, in which participants recall items in whatever order they choose. Existing research has uncovered two main organizational phenomena: ***semantic contiguity effects*** (a tendency for items with similar meanings to be recalled together; Bousfield & Sedgewick, 1944; Jenkins & Russell, 1952; Romney, Brewer, & Batchelder, 1993) and ***temporal contiguity effects*** (a tendency for items studied close in time to be recalled together; Kahana, 1996; Kahana, Howard, & Polyn, 2008).

Semantic contiguity effects can be explained in terms of participants using features of a just-recalled item as a cue for recalling other items (e.g., if you recall a fruit, you can use retrieved fruit features as a cue to recall another fruit). Temporal contiguity effects require a more complex explanation. Modern ***temporal context*** theories (e.g., Howard & Kahana, 2002) posit that temporal contiguity arises because, at encoding, item representations are linked to a slowly changing “context” representation. When an item is recalled, it retrieves the context representation that it was linked to at study, which in turn cues retrieval of items that were studied in similar contextual states. Because (by hypothesis) context changes slowly over time, retrieved context preferentially cues items that were studied *close in time* to the just-retrieved item.

Some theories are agnostic about what information is contained in this context representation and what causes it to drift (e.g., Estes, 1955 and Mensink & Raaijmakers, 1988 both posit random drift). However, more recently, theories like the Temporal Context Model (TCM; Howard & Kahana, 2002) and the Context Maintenance and Retrieval model (CMR; Polyn, Norman, & Kahana, 2009) have set forth a more specific account. According to this account, context is composed (at least in part) of lingering information about recently studied items, which are linked to the memory representation of the currently studied item. For example, if you switch from talking about football to politics, then these theories posit that (for some period of time) the “fading embers” of your football thoughts will persist in your mind and become linked to your memory for the politics discussion. This view converges with recent neuroscientific data showing that information is represented at multiple time scales in the brain, such that some areas only represent the current focus of attention, whereas other areas integrate over longer time scales (Hasson, Yang, Vallines, Heeger, & Rubin, 2008); it also converges with neurophysiological data on “time cells”, showing that different populations of neurons are involved in representing a stimulus, as a function of how long ago the stimulus was presented (e.g., Macdonald, Lepage, Eden, & Eichenbaum, 2011; Howard & Eichenbaum, 2013). In essence, models like TCM and CMR posit that neural populations that represent *preceding* stimulus information get linked to the neural populations that represent *current* stimulus information, thereby contextualizing that information.

The signature prediction of this theory is that, if the lingering thoughts active during an experience X are similar to the lingering thoughts active during an experience Y, then the memories of X and Y should show an elevated probability of being recalled together, because they will have been linked to similar (lingering) information. Remarkably, despite this prediction’s centrality, it has not yet (to our knowledge) been tested. In this experiment, we sought to test this prediction by using multi-voxel pattern analysis (MVPA) of fMRI data (Lewis-Peacock & Norman, 2014; Norman, Polyn, Detre, & Haxby, 2006) to track evidence, at the time of an item’s encoding, for neural representation of the preceding item’s category.

We collected two datasets (n=17 and n=24), following the same data collection procedures for both.[4] Using a multi-voxel classifier of fMRI data that was designed to pick up on lingering traces of the preceding-category, we show that "lingering thoughts" about preceding stimuli influenced the organization of recalls, as predicted by theories of temporal context like TCM – that is, memories encoded with similar "lingering thoughts" about the category of preceding items were more likely to later be recalled close together in time.

## 5.2 Methods

### Participants

For the first dataset, we recruited 17 participants (aged 18-33 years, 11 female) from the Princeton University community. This number of participants was determined a priori. Prior studies using similar paradigms have used on the order of 16 participants (e.g., Polyn et al., 2005, ran 14 participants). We scheduled 17 participants with the goal of getting usable data from 16 participants; it turned out that all 17 participants provided usable data. For the second dataset, we recruited 24 participants (aged 18-29 years, 18 female).

[4] The second dataset was originally collected to replicate a result found in the first dataset. After collection of both datasets, we discovered an error in our original analysis, and here use a corrected analysis to analyze both datasets (the original and corrected analyses are described in "Validating the preceding-category classifier" of the Methods). In this chapter, all results are shown for the individual datasets as well as for the combined data.

All participants provided informed written consent. The study was approved by the Princeton University Institutional Review Board.

## Task

While undergoing functional magnetic resonance imaging (fMRI), participants studied items from different categories. They then performed a recall-by-category task, where participants recalled items from a category that we specified. We used a category structure that allowed us to test how lingering thoughts about preceding items at study affected recall organization at test.

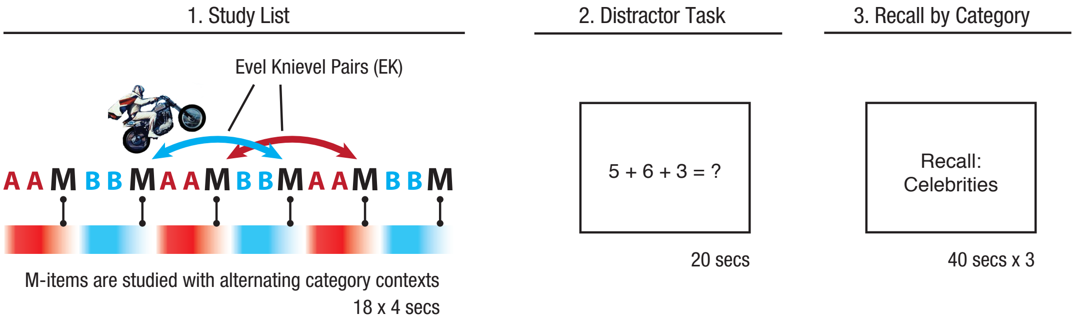


**Figure 1. Trial structure for the recall-by-category task.** Each trial begins with a study list. 18 items were shown one at a time, every 4 seconds. Each study list was composed of items from three different categories (labeled A, B, and M), and the lists were structured as shown. After the study list, participants performed 20 seconds of a distractor task, followed by recall of the items in the M-category (in this example: celebrities), followed next by recall of items in the A and B categories. Transitions at recall between M-items that were studied with the same preceding category are called Evel Knievel (EK) transitions, because they “jump over” temporally nearer M-items (EK transitions may be of length 2 or 4, and may be backwards or forwards).

The structure of the task is shown in Figure 1. At the start of each study-test block, participants were presented with a list of 18 items, one at a time. The items belonged to one of three categories: we schematically refer to them as **A**, **B**, and **M** (where M stands for “main”, because these were the main items of interest; the A- and B-items served to contextualize the M-items, as described below). In any given list, the roles of A, B, and M were mapped one-to-one onto the following three categories of pictures: celebrities, landmarks, and objects. For example, in one list, the A-items might be landmarks, the B-items might be objects, and the M-items might be celebrities. The assignment of categories (celebrity, landmark, object) to roles (A, B, M) was counterbalanced such that — across lists — each category served equally often in the A, B, and M roles.

A new item appeared every 4 seconds, coinciding with the onset of an fMRI image acquisition (each item was shown for 3400 ms, with 600 ms of fixation after each item). Each stimulus presentation was composed of a photograph of a celebrity face, a famous landmark, or a common object, and also the item’s name (e.g., “Eiffel Tower”) presented in text below the photograph; the stimuli were adapted from those used in Morton et al. (2013). To encourage encoding of the items, participants were required to make a category-specific judgment of each item on a 4-point scale. For celebrities, participants were asked to judge, “How much do you love or hate this person?”; for landmarks, they were asked to judge, “How much would you like to visit this place?”; for objects, they were asked to judge, “How often do you come across this object in your daily life?” (Polyn et al., 2005).

After the presentation of the 18 list items, participants performed 20 seconds of a distractor task (self-paced arithmetic problems – summing three random digits, multiple-choice with four choices).

After the distractor task, participants were asked to verbally recall as many items from the list as possible, one category at a time; within a category, participants were allowed to recall freely (i.e., in any order). Participants were first asked to recall M-items (“main items”), and then the A- and B-items; participants were given 40 seconds to recall each category. We analyzed only the recall data from the M-items, but we asked participants to recall the A- and B-items as well, to ensure that they paid attention to those items during study.

There were 12 study-test blocks in total. The experiment task was coded using Psychtoolbox 3 (http://psychtoolbox.org). The verbal recalls for the M-items were annotated using Penn TotalRecall (http://memory.psych.upenn.edu/TotalRecall).

The primary dependent measure of interest was data on the order in which M-items were recalled, as manifested in patterns of *recall transitions*. We say a “transition” has occurred from item X to item Y when participants recall items X and then Y in immediate succession (i.e., without recalling any intervening items).

The key to our study-list structure was that the M-items were preceded by “context items” that alternated in category (A then B then A then B…). According to temporal context theories like TCM, the M-items should be linked to lingering thoughts about the preceding category (either A or B), and this linking to the preceding category should influence the organization of recall. In the absence of this influence, temporal contiguity effects should dominate the patterns of recall, favoring recall transitions between neighboring M-items, as has been previously observed for free recall experiments using study-lists without the same type of alternating semantic structure (e.g., Kahana, 1996; Polyn, Erlikhman, & Kahana, 2011). However, if lingering category information is indeed “contextualizing” M-items in our study-lists, there should be a boost in transition probability between M-items that were preceded by the same category. Because the A and B context items alternated in category, these transitions between M-items with matching “preceding-category context” involve “leaping over” a temporally nearer M-item in favor of a farther M-item; accordingly, we call these transitions “Evel Knievels” (or *EK transitions*), after the daredevil stuntman famous for his motorcycle jumps across canyons, piled cars, and other obstacles. EK transitions could be of length 2 or 4 (jumping over 1 or 3 M-items), in the forwards or backwards directions.

**Overview of fMRI analysis**

As noted above, our main hypothesis was that lingering thoughts relating to preceding items would become linked to M-items at study, thereby resulting in an elevated probability of transitions between M-items that were preceded by the same “context” category (i.e., EK transitions). Importantly, we also expected there to be moment-to-moment variability in the extent to which preceding-category information was represented in participants’ brains; we only expected to see a boost in EK transitions for the subset of trials where preceding-category information actually persisted in participants’ brains. To test this prediction, we used fMRI pattern classifiers to track participants’ thoughts about the preceding category (Lewis-Peacock & Norman, 2014; Rissman & Wagner, 2012). By estimating the level of lingering category information associated with particular M-items, we could make predictions about the order in which these M-items would later be recalled. Specifically, we predicted that — for a pair of M-items X and Y that were preceded by the same context category and thus could later be recalled together as an EK transition — preceding-category information for X and Y (as measured by the classifier) would be more similar when participants actually made the EK transition, compared to when they made a non-EK transition from one of those M-items.

As a control analysis, we used the same logic to address whether the properties of the M-items themselves affected recall order, in our data. Previous work suggests that, on the recall-by-category task, participants might also use information about the semantic category of the items themselves (in addition to retrieved context information) to cue memory recall. If so, the use of current-category cues would lead to clustering-together of M-items that registered neurally as belonging to the same category (Morton et al, 2013). Therefore, we also investigated whether current-category information affected recalls in this way, in our data, in order to isolate our main effect of interest. To measure this potential second effect on recall organization, we used fMRI pattern classifiers to also measure the amount of M-category information elicited by each M-item (we call this current-category information, to distinguish it from preceding-category information). Following the same logic as our main analysis, we investigated whether — for the same pair of M-items X and Y that could later be recalled together as an EK transition — levels of current-category similarity for X and Y were higher when participants actually made the EK transition vs. when they did not.

To show that effects of preceding-category information are distinct from any potential effects of current-category information, we performed a correlation analysis to investigate any correlations between current-category similarity and preceding-category similarity for pairs of items.

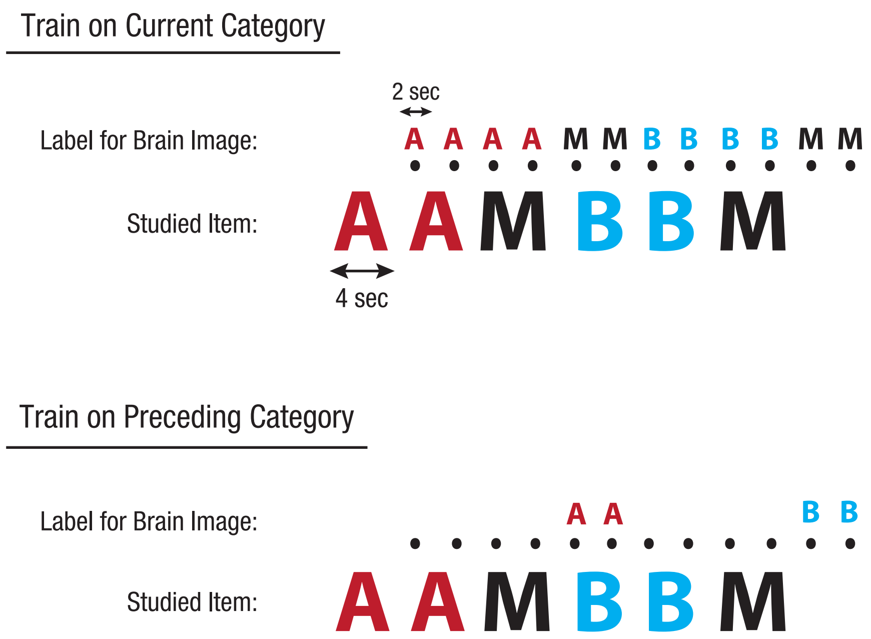


**Figure 2. Labeling of brain images for MVPA classifiers.** We trained and tested MVPA classifiers in two different ways: (1) training and testing on the current semantic category; (2) training and testing on the preceding semantic category. The figure illustrates how brain images (indicated by dots) were labeled for the two classifier types. Brain images were collected every two seconds; stimuli were presented every four seconds (stimulus onset was timed to coincide with the start of an image acquisition). See text for additional differences between our preceding-category classifiers and standard current-category classifiers.

### fMRI acquisition and pre-processing

Functional brain images were acquired using a 3T MRI scanner (Siemens, Skyra) and were preprocessed using FSL (http://fsl.fmrib.ox.ac.uk/fsl/). An echoplanar imaging sequence was used to acquire 40 slices (3mm iso, repetition time (TR) = 2s, echo time (TE) = 30ms, flip angle = 71°). We collected 3 study-test blocks in each scanning run; there were 4 scanning runs in total. The functional images were spatially filtered using a Gaussian kernel (full width at half maximum of 5mm), and then they were temporally filtered using a low-pass cutoff of 0.0077Hz. We performed motion correction using a six-parameter rigid body transformation to co-register functional scans, and then registered the functional scans to an anatomical scan using a 6-parameter affine transformation. Data were spatially normalized by warping each participant's anatomical image to MNI space using a 12-parameter affine transformation. To prepare the data for pattern classification, the activity for each voxel was z-scored within each study-test block.

### MVPA classifier training and testing

Multi-voxel pattern analysis (MVPA) was performed using the Princeton MVPA Toolbox (https://code.google.com/p/princeton-mvpa-toolbox/). We trained two distinct pattern classifiers. First, we trained a classifier to decode information about the category of the current stimulus. Second, we trained a classifier to decode lingering information about the category of the preceding stimuli, based on neural activity from the time of the current stimulus. We trained two distinct classifiers (instead of using just one classifier to decode both current and preceding stimulus identity) because of recent evidence (mentioned above: Hasson et al., 2008; Howard & Eichenbaum, 2013) suggesting that different neural populations may be responsible for coding the current stimulus vs. lingering information about preceding stimuli. The training methods for these two distinct classifier types are described below.

To create training and testing examples for the classifier designed to detect the category of the *current* stimulus, we labeled each brain image with the category of the stimulus presented at that time. Because brain images were acquired every 2 seconds and stimuli were presented every 4 seconds, each stimulus was linked to two brain images. Then we shifted these labels 4 seconds forward in time; this shift accounts for lag in the hemodynamic response measured by fMRI. For example, if the participant studied a celebrity for 4 seconds, then the two images acquired starting 4 seconds and 6 seconds after the onset of the celebrity were labeled as being “celebrity” brain patterns (see Figure 2, top).

To create training and testing examples for the classifier designed to detect the category of the *preceding* stimulus, we took the brain images for which we would expect the peak response to each *M-item* (the same brain images that we used to train a classifier on the *current* category, acquired 4 and 6 seconds after the onset of the M-item), and — instead of labeling those images with the category of the M-item (as we did above) — we labeled those images with the category that *preceded* that M-item (e.g., if the M-item was a celebrity that was preceded by landmarks, we would label those images as being “landmark” brain patterns; see Figure 2, bottom). All other (unlabeled) images were left out of classifier training and testing.

For each participant, we trained three separate preceding-category classifiers – one classifier for the lists where the M-category was celebrities, one for the lists where the M-category was landmarks, and one for the lists where the M-category was objects. In this way, the classifiers could not use current-category information to aid in classifying the preceding category, since the current category was held constant for all training (and testing) examples. To further aid the classifier in focusing on preceding-category information, we used feature selection that selected against voxels that varied significantly with the current category (ANOVA-based feature selection with a threshold of $p$ = 0.05). The next section describes in more detail the rationale for the design of the preceding-category classifier.

For both current-category and preceding-category classifiers, we used logistic regression with L2 regularization (using a regularization penalty of 1; classifier performance was not very sensitive to this parameter). Specifically, we trained a logistic regression classifier for each category to respond with a “1” when an image was labeled with that category and with a “0” when an image was not labeled with that category. Once trained and presented with new input data, these category-specific classifiers output a real value from zero to 1, indicating the degree of neural evidence for the category that it was trained to detect. Classifiers were always trained and tested in a leave-one-block-out fashion — e.g., to apply the classifier to a time point from study-test block 1, the classifier was trained on blocks 2 through 12.

### Validating the preceding-category classifier

Our initial procedure for training a preceding-category classifier (originally applied to the first dataset) produced classifiers that in fact opportunistically used current-category to aid in that classification. Here, we describe the corrected procedure we used to create an improved preceding-category classifier, and we show why it is superior to the more straightforward approach that we originally used.

A standard current-category classifier would be trained by labeling each timepoint (TR) with the category of the current item, after accounting for hemodynamic lag. The most straightforward labeling procedure for a preceding-category classifier would be to label the timepoints of interest with the category of the *preceding* category instead (see Figure 2, bottom). However, we initially did not take any measures to hold the current category constant across lists, so that classifiers trained on these labels could in fact opportunistically use current-category information to aid in the classification of the preceding-category -- information about the current category informs the classifier about what the preceding category is *not*. This negative weighting against the current classifier is visible when we applied the classifier in timecourses of classifier output (Figure 3b). It is especially apparent in the first few timepoints of the study list – these timepoints are not preceded by any A, B, or M items, and so we should expect a true preceding-category classifier to be at chance. However, the classifier knows that the preceding-category can't be A, and shows a clear negative bias against A.

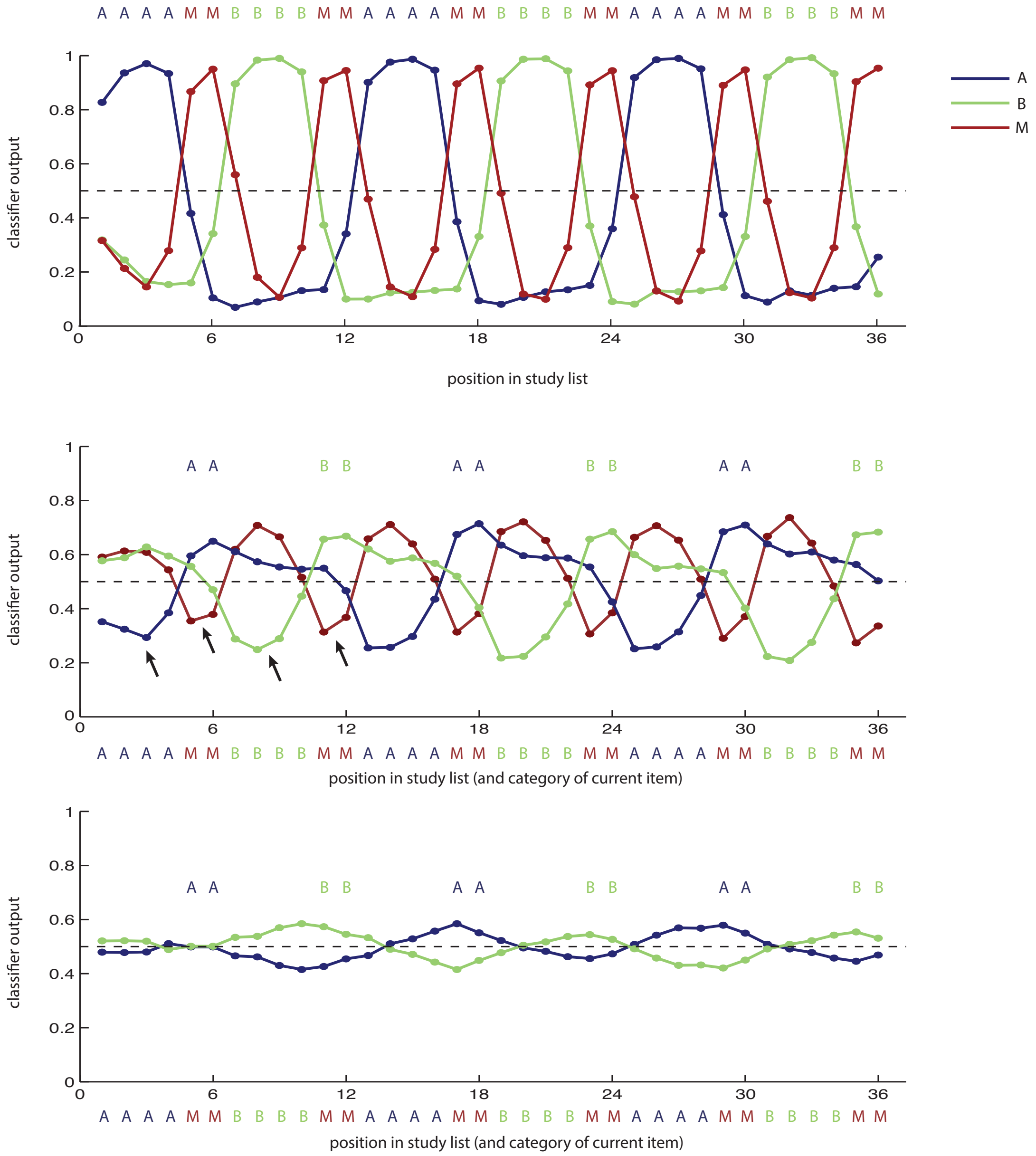


**Figure 3. Timecourses of output for the MVPA logistic regression classifiers, averaged across lists and participants.** Colored letters above each plot indicate the training labels. Colored letters below each plot indicate the category of the current stimulus (after correcting for hemodynamic lag). (a) Outputs for classifiers trained to identify the current category. (b) Outputs for classifiers trained on M-timepoints to identify the preceding category (applied to all timepoints of the list). This version is trained on all lists together. These classifiers show bias against the current-category (black arrows indicate a few examples of negative activation of the current category). (c) Outputs for classifiers trained on M-timepoints to identify the preceding category (applied to all timepoints of the list). *(caption continued on next page)*

*(caption continued from last page)* This version is trained separately for lists with the M-category as celebrity, landmarks, and objects (so that the current category is held constant for all training examples), and used feature selection against voxels with strong current-category information.  These classifiers no longer show a consistent bias against the current-category. They also show a gradual buildup of the A and B categories that peaks at the first TR of each M-item.

To remedy this problem, we made three changes to the classifier. Firstly, we trained three separate classifiers for each participant: one classifier for the lists where the M-category was celebrities, one where the M-category was landmarks, and one where the M-category was objects. In this way, information about the current category was not available to the classifier, since the current-category was held constant for all training (and testing) examples for each classifier (remember that we only used the M-timepoints for preceding-category classification). Secondly, we used a whole-brain mask instead of a temporal-occipital mask, to give the classifier the opportunity to draw from more anterior parts of the brain, if persistent information about recent stimuli is represented there (previous research has shown that this does appear to be the case, e.g. Hasson et al, 2008). Lastly, we implemented feature selection that selected *against* voxels that varied significantly with the current category (we removed these voxels from consideration, using ANOVA-based feature selection with a threshold of $p$ = 0.05).

This classifier training procedure is disadvantaged in that it only has 3 lists for each cross-validated training iteration (rather than 11), and may suffer from having less data. However, as can be seen in Figure 3c, this new version of the classifier has a very different profile from the one in Figure 3b, and no longer shows the same bias against the current category. In fact, as we would expect, these classifiers generally show outputs that slowly ramp up through each block of A- or B-items, peaking at the 1st TR for each M-item.

### Using classifier evidence to compute current-category and preceding-category similarity for pairs of items

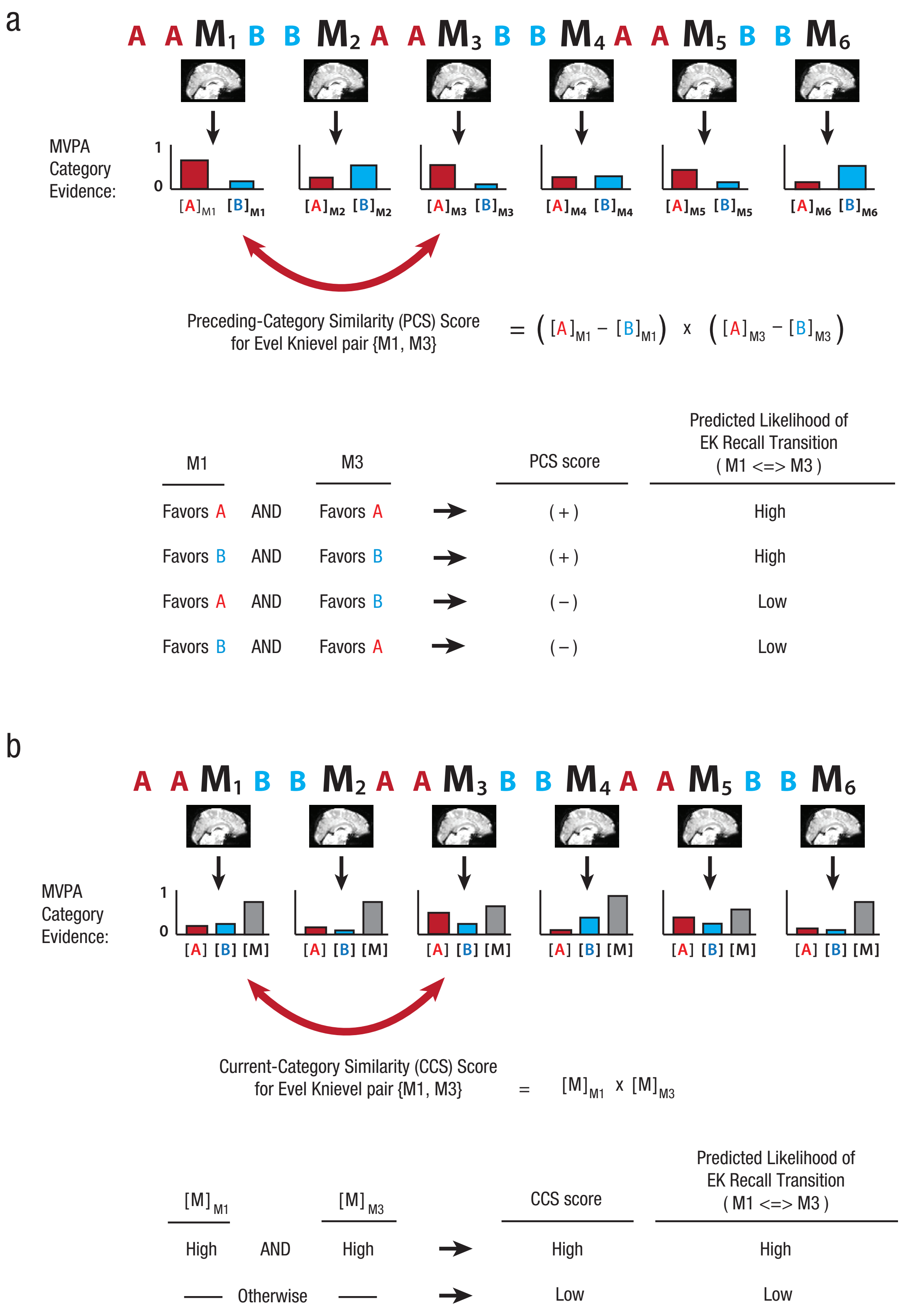
We predicted that, if two M-items were studied with similar profiles of “preceding category information”, participants would be more likely to transition directly between these items at recall (this directly tests the hypothesis that preceding-category information contextualizes the M-items).

To evaluate this prediction, we computed the “preceding category similarity” for each pair of M-items that could potentially form an EK transition at recall (i.e., any pair of M-items preceded by the same “category context”). The preceding-category similarity (***PCS***) measured how much the two M-items registered as being preceded by the same category context.

Preceding-category similarity (PCS) for a potential EK pair was computed as:

$$( [A]_1 - [B]_1 ) \times ( [A]_2 - [B]_2 )$$

where $[A]_1$ is the level of A-category evidence at the time of studying the 1st M-item in the pair, $[A]_2$ is the level of A-category evidence at the time of studying the 2nd M-item in the pair, and so on. Importantly, A-category and B-category evidence in this score was read out using classifiers trained to detect the *preceding* category, described above. The subtractions $[A]_1 - [B]_1$



Preceding-Category Similarity (PCS) Score for Evel Knievel pair {M1, M3} $= \left( [A]_{M1} - [B]_{M1} \right) \times \left( [A]_{M3} - [B]_{M3} \right)$

| M1 | | M3 | | PCS score | Predicted Likelihood of EK Recall Transition ( M1 <=> M3 ) |
|---|---|---|---|---|---|
| Favors A | AND | Favors A | → | ( + ) | High |
| Favors B | AND | Favors B | → | ( + ) | High |
| Favors A | AND | Favors B | → | ( − ) | Low |
| Favors B | AND | Favors A | → | ( − ) | Low |

Current-Category Similarity (CCS) Score for Evel Knievel pair {M1, M3} $= [M]_{M1} \times [M]_{M3}$

| $[M]_{M1}$ | | $[M]_{M3}$ | | CCS score | Predicted Likelihood of EK Recall Transition ( M1 <=> M3 ) |
|---|---|---|---|---|---|
| High | AND | High | → | High | High |
| —— | Otherwise | —— | → | Low | Low |

**Figure 4. Procedure for computing preceding-category and current-category similarity scores.** Caption on next page.

**Figure 4. Procedure for computing preceding-category and current-category similarity scores.** A) MVPA analysis to compute “preceding-category similarity score” (PCS score) for potential EK transitions. Classifiers were trained to identify the preceding category. Classifier outputs were interpreted as levels of evidence for each category. For a given pair of M-items, outputs from these classifiers were combined to form the PCS score, which was designed to measure similarity in lingering levels of the preceding category. B) MVPA analysis to compute “current-category similarity score” (CCS score) for potential EK transitions. Classifiers were trained to identify the current category. For a given pair of M-items, these classifier outputs were multiplied to obtain the CCS similarity score, which was designed to measure similarity in levels of the current category.

and $[A]_2$ - $[B]_2$ measure the “balance” of lingering category evidence (in favor of A vs. B) for the 1st and 2nd M-items. If both M-items strongly favor the same preceding category (both favor A or both favor B, i.e. [A] - [B] for both M-items is strongly positive or strongly negative), then this similarity score is strongly positive (close to +1). In such cases, we would expect a high probability of recall transition between the two M-items, because the MVPA decoders indicate that the M-items were encoded with similar preceding-category contexts. If the M-items strongly favor opposite categories (one favors A and one favors B), then this similarity score is strongly negative (close to -1). In such cases, we would expect a low probability of recall transition between the two M-items (Figure 4a).

As a control analysis, we also performed a parallel analysis to evaluate the degree to which participants were more likely to recall items together if those items triggered similar neural activity corresponding to the *current* category (i.e., basic semantic clustering). Current-category similarity (CCS) for a potential EK pair was computed as:

$$[M]_1 \times [M]_2$$

where $[M]_1$ is the level of M-category evidence associated with the 1st item in the pair, and $[M]_2$ is the level of M-category evidence associated with the 2nd item in the pair. Importantly, M-category evidence in this score was read out using classifiers trained to detect the *current* category. Previous work suggests that the more strongly both M-items favor the (correct) M-category, the greater the “current-category similarity” and the higher the probability that participants should transition between these items (Figure 4b).

**Relating classifier evidence to recall order**

To test our predictions about how recall order depends on preceding-category similarity, we looked at recall of M-items, and separated the observed recall transitions into EK and non-EK transitions. Our goal was to assess whether there were reliable differences in preceding-category similarity (PCS) for potential EK pairs when participants “jumped over” a nearer M-item to make the EK transition, vs. when they made a non-EK transition to the just-nearer M-item. We predicted that PCS (for a pair of M-items that formed a potential EK transition) would be relatively higher when participants actually made the EK transition during recall (vs. when they instead made a non-EK transition to the just-nearer M-item).

We also performed a parallel analysis using current-category similarity (CCS) instead of preceding-category similarity (PCS), to evaluate any effects of current-category similarity on recall organization.

To ensure that we carried out a fair comparison between EK and non-EK transitions (Figure 5), we only analyzed EK transitions where it was actually possible for participants to instead have made a non-EK transition in the same direction, to the just-nearer M-item (i.e., the just-nearer M-item had not already been recalled). Likewise, we only analyzed non-EK transitions where it was actually possible for participants to instead have made an EK transition in the same direction, to the just-farther M-item — we excluded non-EK transitions where the just-farther M-item had already been recalled, and we also excluded non-EK transitions where participants transitioned backward to the first M-item or forward to the last M-item on the list (in these cases, there *was* no just-farther M-item). Because of this extra exclusion condition for non-EK transitions, we ended up excluding more non-EK transitions than EK transitions: on average, we excluded 17% of EK transitions (95% CI: 13-22%) and 46% of non-EK transitions (95% CI: 41-51%).

In order to capture the relative strength in preceding- (or current-) category similarity for a potential EK pair, compared to its just-nearer potential nonEK pair, we computed PCS (or CCS) for both pairs of items and took the difference between the two scores.

### Statistics and confidence intervals

For all of our analyses looking (separately) at behavioral data or neural data, we computed random-effects bootstrap confidence intervals on the mean by resampling participants with replacement (Efron & Tibshirani, 1986).

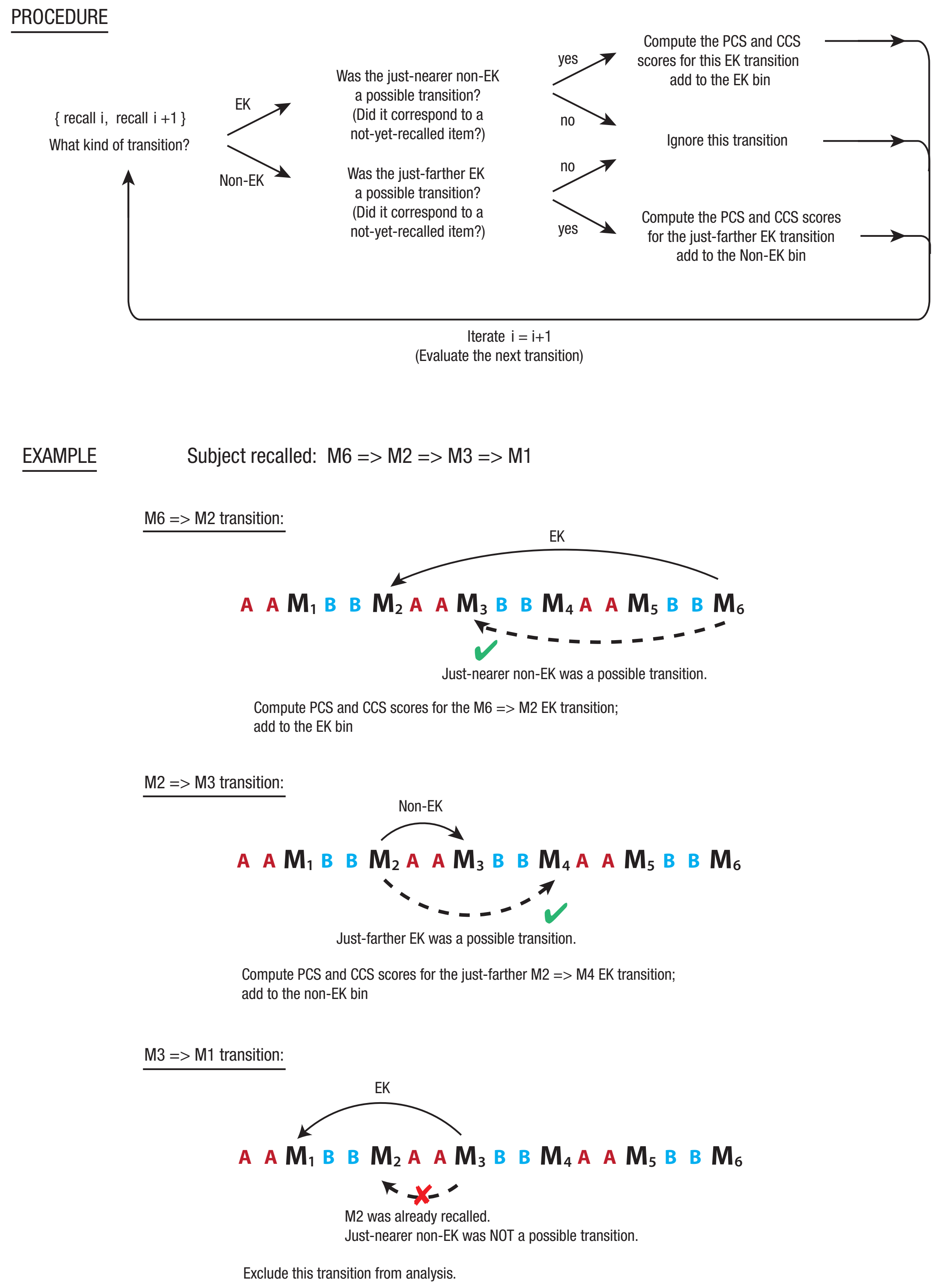

PROCEDURE
{ recall i, recall i +1 }
What kind of transition?
EK
Non-EK
Was the just-nearer non-EK a possible transition? (Did it correspond to a not-yet-recalled item?)
Was the just-farther EK a possible transition? (Did it correspond to a not-yet-recalled item?)
yes
no
no
yes
Compute the PCS and CCS scores for this EK transition add to the EK bin
Ignore this transition
Compute the PCS and CCS scores for the just-farther EK transition add to the Non-EK bin
Iterate i = i+1
(Evaluate the next transition)
EXAMPLE
Subject recalled: M6 => M2 => M3 => M1
M6 => M2 transition:
EK
A A M1 B B M2 A A M3 B B M4 A A M5 B B M6
✓
Just-nearer non-EK was a possible transition.
Compute PCS and CCS scores for the M6 => M2 EK transition; add to the EK bin
M2 => M3 transition:
Non-EK
A A M1 B B M2 A A M3 B B M4 A A M5 B B M6
✓
Just-farther EK was a possible transition.
Compute PCS and CCS scores for the just-farther M2 => M4 EK transition; add to the non-EK bin
M3 => M1 transition:
EK
A A M1 B B M2 A A M3 B B M4 A A M5 B B M6
✗
M2 was already recalled.
Just-nearer non-EK was NOT a possible transition.
Exclude this transition from analysis.

**Figure 5. Procedure for aggregating valid EK and non-EK transitions, to ensure fair comparison between the two transition types.** Caption on next page.

**Figure 5. Procedure for aggregating valid EK and non-EK transitions, to ensure fair comparison between the two transition types.** In our main analysis, we only included EK transitions where a non-EK transition to the just-nearer M-item would also have been possible (i.e., the just-nearer M-item had not already been recalled), and we only included non-EK transitions where an EK transition to the just-farther M-item would also have been possible. If transitions to the just-nearer M-item (for an EK transition) or the just-farther M-item (for a non-EK transition) were not possible, then we ignored this transition and continued to the next. Otherwise, we considered it a valid transition and included it in our analysis. For EK transitions, we computed the similarity score for the EK transition; for non-EK transitions, we computed the similarity score for the just-farther EK transition. The figure shows an example recall sequence (M6, M2, M3, M1) for a particular list; for this sequence, we would include M6=>M2 as a valid EK transition, include M2=>M3 as a valid non-EK transition, and exclude M3=>M1 as an invalid EK transition.

When assessing differences between conditions, we computed bootstrap confidence intervals on the difference between the means. In the text, these are reported as 95% confidence intervals. In the results figures, these bootstrap distributions and confidence intervals are displayed using cat's eye plots.

## 5.3 Results

### Behavioral results

On average, participants correctly recalled 54.8% of the M-items that they studied (95% CI: 51.5–58.1%). Broken down by category, participants recalled 63.4% of celebrity M-items (95% CI: 60.0–66.8%), 60.5% of landmark M-items (95% CI: 55.6% to–65.2%), and 40.4% of object M-items (95% CI: 36.1–44.6%). Participants complied with our instructions not to repeat themselves during free recall (i.e., they never recalled the same M-item twice during a single recall period). Participants occasionally made intrusions (i.e., recalled items not on the current study list); transitions involving intrusions were not included in our EK analysis (e.g., if a participant recalled item M2, an intrusion next, and finally item M4 after that, then neither the M2=>intrusion nor the intrusion=>M4 transitions were included in our analysis). On average, each participant made 0.13 intrusions per list (95% CI: 0.092–0.18). Of these intrusions, 26% on average were the names of items studied on previous lists (95% CI: 14–42%); the other intrusions were names that had not appeared anywhere in the experiment. On average, each subject made 10.4 valid EK transitions (95% CI: 9.0–11.8) and 16.6 valid non-EK transitions (95% CI: 15.0–18.6), where "valid" is as defined above and in Figure 5. These behavioral results are reported in Table 1 for each dataset individually.

| | **Dataset 1 (n=17)** | **Dataset 2 (n=24)** |
|---|---|---|
| % of M-items correctly recalled | 55.7% (51.9 - 59.6%) | 54.2% (49.2 - 59.2%) |
| % of celebrity M-items correctly recalled | 64.2% (59.6 - 69.1%) | 62.8% (58.0 - 67.5%) |
| % of landmark M-items correctly recalled | 62.0% (55.8 - 68.0%) | 59.4% (52.6 - 66.3%) |
| % of object M-items correctly recalled | 40.9% (36.57 - 45.3%) | 40.0% (33.4 - 46.6%) |
| mean number of intrusions per list | 0.132 (0.078 - 0.196) | 0.132 (0.083 - 0.212) |
| % of intrusions that were prior-list items | 29.6% (11.1 - 66.7%) | 23.7% (10.5 - 47.4%) |
| mean # of valid EK transitions | 10.6 (9.0 - 12.6) | 10.3 (8.5 - 12.1) |
| mean # of valid non-EK transitions | 17.1 (15.1 - 18.9) | 16.3 (14.0 - 19.4) |

*Table 1. Behavioral results for each dataset individually (95% confidence intervals in parentheses).*

**Basic classifier results**

Before relating the classifier output to recall behavior, we first wanted to establish that the preceding-category classifier was decoding category identities at above-chance levels.

For the classifiers trained to decode the preceding category, we computed accuracy for each fMRI image based on whether classifier evidence for the correct context category (A or B: whichever one actually preceded this particular M-item) was greater than classifier evidence for the incorrect context category. For this 2-way classification, chance is 50%. The observed level of accuracy was 57% for lists with celebrities as the M-category (95% CI: 54-60%), 57% for lists with landmarks as the M-category (95% CI: 53-61%), and 58% for lists with objects as the M-category (95% CI: 54-61%). (These classifier results are reported in Table 2 for the individual datasets.) Importantly, these accuracy percentages only denote the percentage of outputs that matched the preceding-category *labels* that we provided to the classifier—not the match to the participants' actual neural content. We believe that the output of the classifier in fact reflects a noisy estimate of meaningful fluctuations in the extent to which preceding-category information lingered in participants' brains. In our main analysis, this variability in the classifier output is what allows us to make predictions about when participants will make EK transitions.

| M category | **Dataset 1 (n=17)** | **Dataset 2 (n=24)** |
|---|---|---|
| celebrities | 57% (54 - 59%) | 57% (54 - 60%) |
| landmarks | 58% (55 - 62%) | 56% (53 - 59%) |
| objects | 57% (54 - 60%) | 58% (55 - 60%) |

*Table 2. Basic classifier results for each dataset individually. Reported are mean classifier accuracies (and 95% confidence intervals) for the three different classifiers.*

### Relating classifier evidence to recall order

As predicted, levels of preceding-category similarity (i.e., PCS scores) were higher for a pair of M-items when participants made the EK transition between them, vs. when they instead made a non-EK transition to the just-nearer M-item (Figure 6a, top). That is, participants were more likely to recall two M-items together if the M-items were encoded with similar lingering information about preceding items. This result provides direct support for the idea that lingering thoughts relating to preceding items serve to contextualize memories and organize subsequent recall. The result was also observed for the individual datasets, although the result was not significant for the first dataset on its own, which showed a smaller effect overall (Figure 6a, middle and bottom).

We did not find a corresponding effect of current-category similarity on recall order in the combined dataset – that is, we did not find that levels of current-category similarity (i.e., CCS scores) were significantly higher for a pair of M-items when participants made the EK transition between them vs.

when they instead made a non-EK transition to the just-nearer M-item (Figure 6b, top). When analyzing individual datasets, we did find a current-category effect in the 1st dataset, but not in the 2nd dataset. In fact, in the 2nd dataset, the effect fell marginally in the opposite direction (Figure 6b, middle and bottom). The lack of a robust effect of current-category similarity on suggests that the preceding-category results were distinct from any potential effects of current-category similarity.

However, although current-category similarity in the full dataset did not differ significantly between EK vs non-EK transitions made, it was numerically higher for EK transitions made. Thus, to ensure that our preceding-category results were not driven by any effects of current-category similarity, we computed the correlation between current-category similarity and preceding-category similarity. There was no significant correlation (mean correlation for each participant 0.016; 95% CI: -0.011–0.044).

Effect sizes are reported in Table 3, for effects of preceding-category similarity and current-category similarity on EK transitions.

| | **Combined data**<br>(n=41) | **Dataset 1**<br>(n=17) | **Dataset 2**<br>(n=24) |
|---|---|---|---|
| Difference in PCS scores<br>for EK vs. nonEK transitions made | 0.64 | 0.32 | 0.92 |
| Difference in CCS scores<br>for EK vs. nonEK transitions made | 0.33 | 0.75 | -0.04 |

*Table 3. Effect sizes (reported as Cohen's d) for effect of preceding-category similarity (PCS) scores and current-category similarity (CCS) scores on recall transitions.*

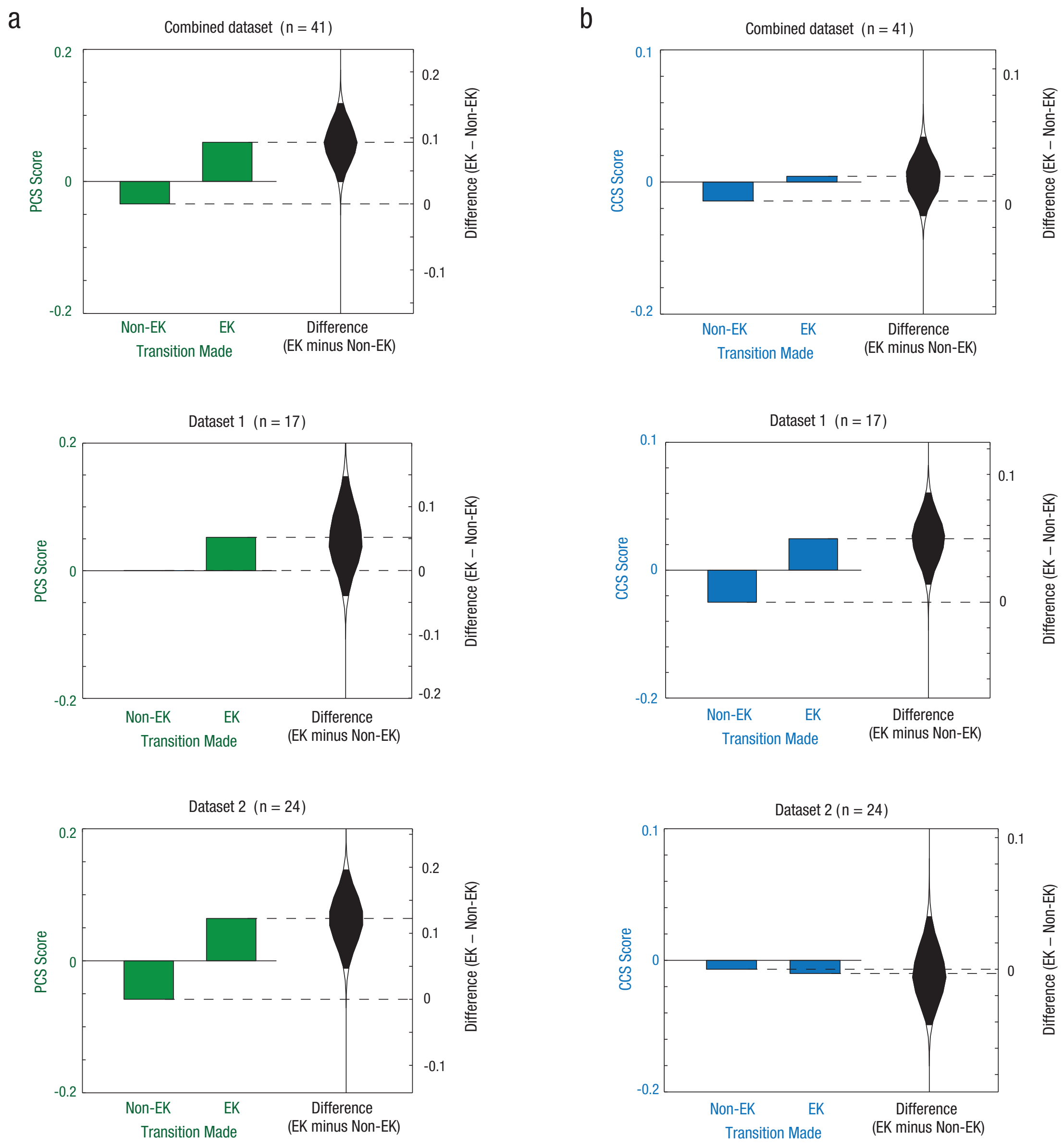


**Figure 6. Main results.** For EK transitions that actually occurred vs. EK transitions that did not occur (because a non-EK transition occurred instead): (a) preceding-category similarity (PCS) scores, (b) current-category similarity (CCS) scores. Cat's-eye plots show bootstrap distributions for the difference in PCS/CCS scores between EK transitions that did occur vs. those that did not. Results are shown for the combined data, and for the individual datasets. Black shaded areas of cat's-eye plots indicate 95% confidence intervals.

## 5.4 Discussion

In this study, we used fMRI pattern classification to track lingering traces of preceding thoughts, and we showed that memories encoded with similar "lingering thoughts" (about the category of preceding items) were more likely to later be recalled together. The idea that items are contextualized by the "fading embers" of recently studied items is a central assertion of extant models of temporal context and memory (e.g., Howard and Kahana 2002). Our results provide the most direct evidence to date in support of this view.

These effects of "lingering thoughts" on recall order are distinct from previously documented effects by information about *current* stimuli. These previous studies include Morton et al. (2013), who found that the degree of category-specific activity elicited by a studied item predicted category clustering on a free recall test (see also Kuhl et al, 2012, who found that category-specific activity at encoding predicted cued recall success at test). In our combined dataset, we did not find that information about the current item's category predicted which items would later be recalled close together in time. Furthermore, we did not find any correlation between measures of preceding-category similarity and current-category similarity. This indicates that information about the *preceding* item's category exerted influences that were not driven by, and were distinct from, any potential influences of the current item's category.

Our results, taken together with previous experiments, point to a synthesis whereby multiple time scales of representation influence recall organization, in distinct ways. Although we did not robustly observe this effect in our data, previous studies have shown that information about what is *currently* happening gets encoded into the memory trace, leading to semantic clustering effects — two events that have similar content activate overlapping populations of neurons, such that thinking of one event automatically cues the other (similar) event (e.g. Morton et al, 2013). Newly, our results show that information about what was *recently* happening is also encoded into the memory trace. Under normal circumstances, this can lead to temporal clustering (if participants see events A, B, C in sequence, lingering information about A gets encoded along with both B and C, leading to enhanced transitions between B and C). In our study, however, we deliberately structured study lists so that encoding of preceding-category information worked *against* temporal clustering — to the extent that participants were integrating preceding-category information into their memory traces, they should make "Evel Knievel" recall transitions that jump over nearer items, which is exactly what we saw.

One limitation of our study is that we only tracked thoughts relating to the *immediately preceding* category. As such, our results do not, on their own, discriminate between dual-store memory models (which posit that recently studied items linger in a short-term memory store, so that adjacent items are directly associated with each other during study; Atkinson & Shiffrin, 1968;

Raiijmakers & Shiffrin, 1980) and memory models like TCM (which posit that items are contextualized by linking them to a running average of recently presented items). It is worth emphasizing that both of these accounts (dual-store models and models like TCM) posit that activity relating to preceding items persists in some form, and is linked to the current item. Thus, the effect of lingering information on recall organization is a key prediction of a wide set of prominent theories of memory, for which our experiment is the most direct test to date.

## 5.5 References

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. The Journal of General Psychology, 30(2), 149–165. doi:10.1080/00221309.1944.10544467

Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate levels of activation lead to forgetting in the think/no-think paradigm. Neuropsychologia, 51(12), 2371–2388. doi:10.1016/j.neuropsychologia.2013.02.017

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1(1), 54–75. doi:10.1214/ss/1177013815

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. Psychological Review, 62(3), 145.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. Journal of Neuroscience, 28(10), 2539–2550. doi:10.1523/JNEUROSCI.5487-07.2008

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425–2430. doi:10.1126/science.1063736

Howard, M. W., & Eichenbaum, H. (2013). The hippocampus, time, and memory across scales. Journal of Experimental Psychology: General, 142(4), 1211–1230. doi:10.1037/a0033621

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. Journal of Mathematical Psychology, 46(3), 269–299. doi:10.1006/jmps.2001.1388

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? Journal of Memory and Language, 46(1), 85–98. doi:10.1006/jmla.2001.2798

Jenkins, J. J., & Russell, W. A. (1952). Associative clustering during recall. The Journal of Abnormal and Social Psychology, 47(4), 818–821. doi:10.1037/h0063149

Kahana, M. J. (1996). Associative retrieval processes in free recall. Memory & Cognition, 24(1), 103–109.

Kahana, M. J., Howard, M. W., & Polyn, S. M. (2008). Associative retrieval processes in episodic memory. In H. L. Roediger III (Ed.), Cognitive psychology of memory (Vols. 1-4, Vol. 2). Oxford: Elsevier. Retrieved from https://memory.psych.upenn.edu/files/pubs/KahaEtal08.pdf

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. Proceedings of the National Academy of Sciences, 111(24), 8997–9002. doi:10.1073/pnas.1319438111

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences, 103(10), 3863–3868. doi:10.1073/pnas.0600244103

Kuhl, B. A., Bainbridge, W. A., & Chun, M. M. (2012). Neural reactivation reveals mechanisms for updating memory. The Journal of Neuroscience, 32(10), 3453–3461. doi:10.1523/JNEUROSCI.5846-11.2012

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. Proceedings of the National Academy of Sciences, 108(14), 5903–5908. doi:10.1073/pnas.1016939108

Kuhl, B. A., Rissman, J., & Wagner, A. D. (2012). Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. Neuropsychologia, 50(4), 458–469. doi:10.1016/j.neuropsychologia.2011.09.002

Lewis-Peacock, J. A., & Norman, K. A. (2014). Competition between items in working memory leads to forgetting. Nature Communications, 5. doi:10.1038/ncomms6768

Lewis-Peacock, J. A., & Norman, K. A. (2014). Multivoxel pattern analysis of fMRI data. In M. Gazzaniga & R. Mangun (Eds.), Cognitive Neurosciences V (pp. 911-920). Cambridge, MA: MIT Press.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal “time cells” bridge the gap in memory for discontiguous events. Neuron, 71(4), 737–749. doi:10.1016/j.neuron.2011.07.012

Manning, J. R., Kahana, M. J., & Norman, K. A. (2014). The role of context in episodic memory. In M. Gazzaniga & R. Mangun (Eds.), Cognitive Neurosciences V (pp. 557-566). Cambridge, MA: MIT Press.

McDuff, S. G. R., Frankel, H. C., & Norman, K. A. (2009). Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. The Journal of Neuroscience, 29(2), 508–516. doi:10.1523/JNEUROSCI.3587-08.2009

Mensink, G.-J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. Psychological Review, 95(4), 434–455. doi:10.1037/0033-295X.95.4.434

Morton, N. W., Kahana, M. J., Rosenberg, E. A., Baltuch, G. H., Litt, B., Sharan, A. D., … Polyn, S. M. (2013). Category-specific neural oscillations predict recall organization during memory search. Cerebral Cortex, 23(10), 2407–2422. doi:10.1093/cercor/bhs229

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences, 10(9), 424–430. doi:10.1016/j.tics.2006.07.005

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. Science, 310(5756), 1963–1966. doi:10.1126/science.1117645

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009a). A context maintenance and retrieval model of organizational processes in free recall. Psychological Review, 116(1), 129–156. doi:10.1037/a0014420

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009b). Task context and organization in free recall. Neuropsychologia, 47(11), 2158–2163. doi:10.1016/j.neuropsychologia.2009.02.013

Polyn, S. M., Erlikhman, G., & Kahana, M. J. (2011). Semantic cuing and the scale insensitivity of recency and contiguity. Journal of Experimental Psychology: Learning, Memory, & Cognition, 37(3), 766-775. Doi:10.1037/a0022475

Poppenk, J., & Norman, K. A. (2014). Briefly cuing memories leads to suppression of their neural representations. The Journal of Neuroscience, 34(23), 8010–8020. doi:10.1523/JNEUROSCI.4584-13.2014

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. Annual Review of Psychology, 63(1), 101–128. doi:10.1146/annurev-psych-120710-100344

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. Psychological Science, 4(1), 28–34. doi:10.1111/j.1467-9280.1993.tb00552.x

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. Neuron, 75(1), 168–179. doi:10.1016/j.neuron.2012.05.010

# 6 General Discussion: Relating “state”, “context”, and “schemas”

The concepts of states, context, and schemas are central to many active areas of research. In recent years, researchers are starting to recognize the importance of more concretely describing their contents and construction (as implemented by humans and animals, as well as on a more abstract theoretical basis), and also elucidating their neural implementation. A primary goal of this thesis was to begin filling in these gaps in our understanding – what kind of information is used to construct a representation of the current situation? How is the inference performed, and how is it realized neurally?

Until now, state, context, and schemas have been, for the most part, considered as three separate ideas. A second goal of this thesis work was to bring together these three ideas, both theoretically and in terms of their neural underpinnings.

## 6.1 Similarities in the theoretical and neural representations of states and schemas

In this work, I have focused on representations of the current situation for two cognitive processes – decision-making and episodic memory. Normatively, the inference procedure and contents of the representation for the two cognitive processes should be strongly related, though not necessarily identical. For decision-making, a representation of the current situation (the “state”) should include information that is relevant for taking the right actions at the current time. In other words, the representation should capture the contingencies between actions and outcomes. In contrast, the purpose of episodic memory is to store information that could be useful for action and decision-making at a *future* time, and it is optimal to *label* that information in such a way that the relevant information will be retrieved at that future time. Due to the associative nature of memory, this is accomplished by labeling the memories with concepts that are likely to be reactivated when the same action-outcome contingencies are encountered again (because those concepts will in turn activate the relevant memories). This can be accomplished in a straightforward way by directly labeling a memory with one’s inference about the current situation – in other words, this label would consist of a representation of state that is similar to the representation useful for decision making at the *current* time.

Given the similarity of the desired representations and inference procedure of situation for memory and decision-making, one parsimonious solution is for the two cognitive processes to draw on the *same neural circuits* for processing the current situation. Previous research (see Section 2.1) has indeed suggested a similar locus of representation within the brain for states and schemas – the overlapping areas of ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC). In Chapter 4, I showed that, when the underlying situation is not directly observable, representations in this vmPFC/OFC region seem to take the form of a probability distribution over possible situations (as opposed to, say, the best guess of the situation).

For the common case of decision-making under uncertainty about the current state, it is well established that this probability distribution over possible states (the *belief state*) can serve as a stand-in Markov representation whose values can be updated from experience, using the standard machinery of reinforcement learning (Kaelbling et al, 1998). Thus, it makes sense that a brain area implicated in the representation of task state would in fact contain representations of a *probability distribution* over states. Further studies are required to reinforce the evidence presented here that vmPFC/OFC represents belief states used for reinforcement learning. For example, it is not clear what the relationships are between posterior-probability-related activity patterns in vmPFC/OFC and those in areas known to be involved further downstream in decision making, such as parietal cortex.

## 6.2 Organization of memory by situation representations

In contrast to belief states in reinforcement learning, it is a new idea that the schemas used to organize memory might also take the form of a probability distribution over the possible underlying situations. On the one hand, it may make sense for memory labels to take this form; that is, memories may be associated with different situation labels in accordance with how likely that situation underlies the current events (which are being entered into memory) – then, in the future, memories can be reactivated in accordance with how strongly they are linked to a representation of the situation encountered at the future time.[5] At the same time, it is also possible that a probability distribution in vmPFC/OFC serves as an intermediate computation in the inference of schemas, and that memories are not organized or incorporated into the representations of the probability distributions themselves. This remains an open question.

Our results in Chapter 5 may appear to support the idea of an alternate organizing principle for our memories, by supporting existing theories that suggest that memories are labeled with the semantics of recent experience. In addition, previous work has indicated that memories are also frequently organized according to spatial information, time, and emotional state. However, what these types of information (usually called “context”) have in common is that they tend to be very diagnostic of the current situation, and so can serve as useful heuristic proxies for the underlying situation.

There remain a few important open questions that might be addressed in the near term. The work presented in this thesis has shown evidence that an area implicated in the representation of schemas also shows

[5] Analogous to belief states for decision-making, a possibly equivalent formulation of this is to state that memories should be labeled with a single situation label that is in fact a probability distribution over situations, and should be reactivated in accordance with the similarity of the label with the current probability distribution. For example, memory X is labeled with 100% certainty of situation A, memory Y is labeled with 50% situation A and 50% situation B, and memory Z is labeled with 100% certainty of situation B. If, at the current time, we infer that we are in situation A with 90% probability, then (according to this idea) we would be most likely to reactivate memory X and least likely to reactivate memory Z.

representation of a posterior probability distribution over the latent causes that underlie our observations. It has *not* shown that this posterior probability distribution actually organizes and affects memories in the way that schemas are expected to do. Even if this posterior probability distribution is shown to act as a schema or like context for memories, it may not be the only organizing signal for memories. Although certain streams of information known to organize memories (e.g. information about space, time, emotional state, or preceding stimuli, as we showed in Chapter 5) tend to be very diagnostic of the underlying situation, they may not actually be processed through an inference of the situation before being used to organize memories. To answer whether they are or not, we may wish to show whether or not these other streams of information affect the organization of memory even when they are not particularly diagnostic of the underlying situation, and also to see whether or not their use tends to activate parts of the brain that we find to be involved in inference or representation of the current situation.

## 6.3 Conclusion

There remain many open and important questions regarding the inference and representation of situation, and the way these representations are used by other cognitive processes. We have only begun to map out the neural circuits involved in the representation of situation, and research investigating the inference of situation (including neural implementations) is similarly in its infancy, but shows a lot of promise for progress in coming years.