Article

Neuron

Human Orbitofrontal Cortex Represents a Cognitive Map of State Space

Highlights

- We tested a novel theory of OFC function directly in humans with fMRI
- Multivariate pattern analysis showed evidence for state encoding in OFC
- Performance within and across participants was related to state encoding in OFC
- The findings provide strong support for the state representation theory of OFC

Authors

Nicolas W. Schuck, Ming Bo Cai, Robert C. Wilson, Yael Niv

Correspondence

nschuck@princeton.edu

In Brief

Schuck et al. present evidence that orbitofrontal cortex contains an up-to-date representation of task-related information during decision making. This "state" representation might provide important input for efficient reinforcement learning and decision making elsewhere in the brain.

Schuck et al., 2016, Neuron *91*, 1402–1412
September 21, 2016
http://dx.doi.org/10.1016/j.neuron.2016.08.019

CellPress

CellPress

Neuron
Article

Human Orbitofrontal Cortex Represents a Cognitive Map of State Space

Nicolas W. Schuck,[1,3,*] Ming Bo Cai,[1] Robert C. Wilson,[2] and Yael Niv[1]
[1]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Washington Road, Princeton, NJ 08544, USA
[2]Department of Psychology, University of Arizona, 1503 East University Boulevard, Tucson, AZ 85721, USA
[3]Lead Contact
*Correspondence: nschuck@princeton.edu
http://dx.doi.org/10.1016/j.neuron.2016.08.019

SUMMARY

Although the orbitofrontal cortex (OFC) has been studied intensely for decades, its precise functions have remained elusive. We recently hypothesized that the OFC contains a "cognitive map" of task space in which the current state of the task is represented, and this representation is especially critical for behavior when states are unobservable from sensory input. To test this idea, we apply pattern-classification techniques to neuroimaging data from humans performing a decision-making task with 16 states. We show that unobservable task states can be decoded from activity in OFC, and decoding accuracy is related to task performance and the occurrence of individual behavioral errors. Moreover, similarity between the neural representations of consecutive states correlates with behavioral accuracy in corresponding state transitions. These results support the idea that OFC represents a cognitive map of task space and establish the feasibility of decoding state representations in humans using non-invasive neuroimaging.

INTRODUCTION

Imagine deciding between Apple and IBM. Investing in the stock market, you might consider which of the two companies is more financially promising. Shopping for a new laptop, on the other hand, you might rather evaluate the price and quality of display of the different machines and not the portfolios of their manufacturers. The neural mechanisms that evaluate choices, therefore, require the same options to be represented differently in the brain when engaged in different tasks, making the correct representation of the environment an important prerequisite for sound decision-making (Dayan 1993; Sutton and Barto, 1998).

In the computational framework of reinforcement learning (RL; Sutton and Barto, 1998), the collection of information that is relevant to a given decision is called the "state," and decision-making proceeds by comparing the values of different actions at each state. Much research has investigated the location and nature of neural value representations (Bartra et al., 2013; Chase et al., 2015). But where in the brain are the corresponding task-specific states represented?

The capability to flexibly represent currently relevant information is widely believed to reside in prefrontal cortex (Duncan 2001). Of specific interest, the orbitofrontal cortex (OFC) has been implicated in various decision-making functions (Stalnaker et al., 2015) and is known to have particularly wide-ranging connectivity to sensory areas of all modalities, as well as to cortical and subcortical areas related to memory, learning, and attention (Cavada et al., 2000; Kringelbach and Rolls, 2004). Different theories regarding the nature of the information represented in the OFC have suggested, for instance, that either economic values (Padoa-Schioppa and Assad, 2006), emotions (Bechara et al., 2000), or cue-outcome associations (Kringelbach, 2005) are represented and used to inform the decision-making process. But neurophysiological studies have also shown that OFC neurons can represent specific aspects of the current decision-making problem, such as sensory properties of outcomes (McDannald et al., 2014) or integrated schemas detailing the context, position, and reward associated with objects (Farovik et al., 2015). Unifying these accounts and data, we have recently suggested that in order to provide a decision-relevant summary of the environment, OFC may flexibly represent different quantities depending on the task at hand (Wilson et al., 2014). Specifically, we hypothesized that these changing signals in OFC effectively represent the current state within the task, in a way that is tailored for the operation of RL-based decision-making mechanisms in the basal ganglia (Niv 2009). Moreover, because OFC lesions lead to impairments particularly in tasks that involve states that are difficult to distinguish based on sensory input alone (called "partially observable states" or "hidden states") (Wilson et al., 2014; Bradfield et al., 2015), we proposed that OFC's critical contribution to decision-making is to represent such hidden states. Fully observable states may also be represented in OFC, however, these representations appear less critical for decision-making. This hypothesis makes specific predictions regarding information encoding in the OFC that go beyond a general increase in OFC activity in certain tasks but not others. We therefore focused on analyzing multivoxel activity patterns in the OFC, to test whether information about the current hidden task state can be found in this area.

RESULTS

In order to test this theory in humans, we developed a task in which participants' decisions depended on hidden information

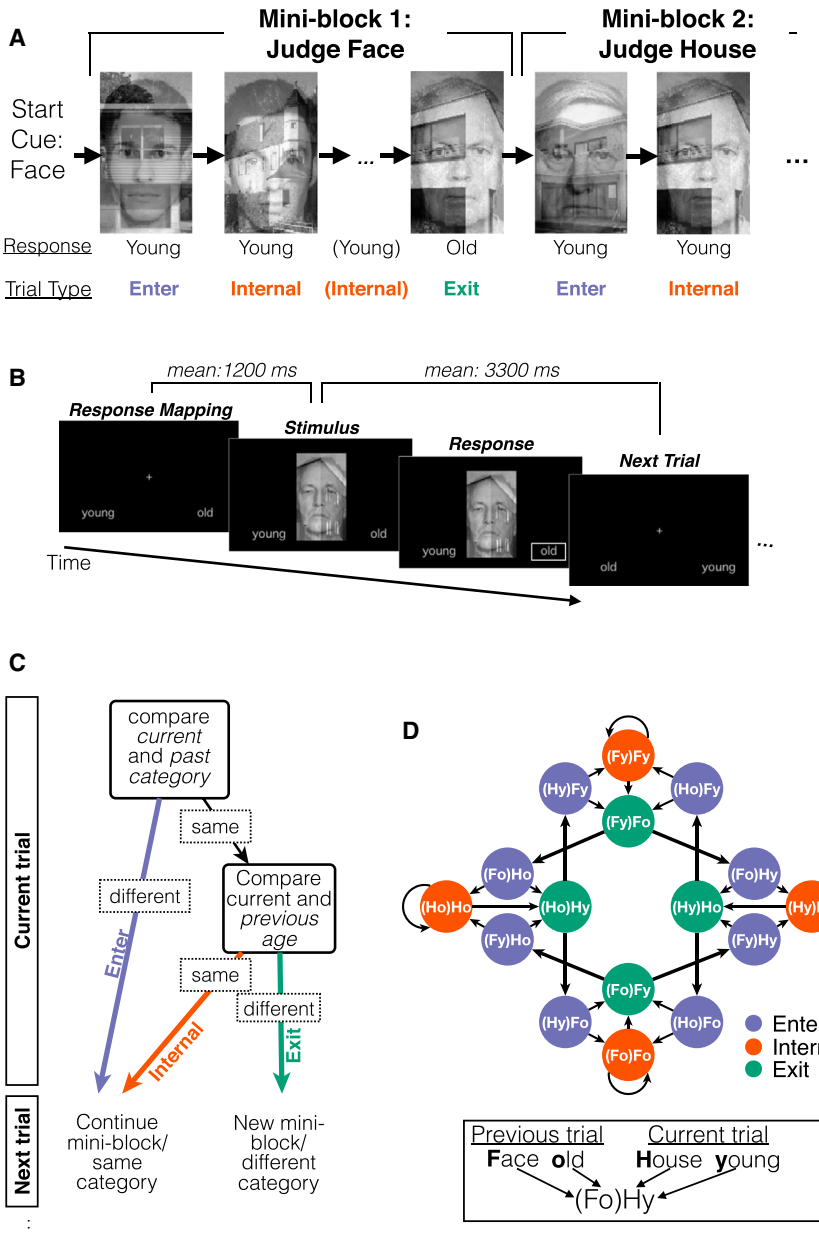

Figure 1. Experimental Task

(A) Example of trial sequence in the task. Participants began by judging the age (young versus old) of the cued category (face or house). In the following trials, they continued to judge the age of the same category until an age change occurred. On the trial following an age change, participants had to switch to judging the other category, with the first trial after the change determining the new age of that category to which subsequent trials needed to be compared. These rules created an alternating mini-block structure of judging either the age of faces or houses. The first trial after a category switch was the trial in which the mini-block was entered (Enter trial), trials in which the category and age repeated are denoted Internal trials, and the trial in which the age changed was the end of the current mini-block (Exit trial).

(B) Trial structure. Each trial started with a fixation cross and the display of the randomly determined response mapping for the current trial. Following the stimulus display, participants had up to 2,750 ms to make their response before the next trial started, while the stimulus duration was independent of the response time and lasted on average 3,300 ms. Responses were followed by a box around the chosen option. Wrong responses led to a repetition of the same (Enter trials) or preceding trial (Internal and Exit trials), accompanied by a written reminder of the current category.

(C) Mental operations involved in different trial types. The diagram illustrates how currently hidden information about the previous age as well as the current and previous category must be factored into the decision-making process.

(D) Possible transitions between states during the task. Each circle denotes a particular state (see legend). Arrows indicate possible transitions and node colors indicate the trial types.

involving memory of past events and knowledge about the current phase of the task. Using past events to contextualize current decisions and values is a common aspect of real-world tasks, such as brokering on the stock market where the overall trend of a stock is more important than its current absolute value. Specifically, in our task participants were asked to judge the age (old versus young) of either a face or a house. The presented images always contained both a face and a house spatially superimposed (Figure 1A), but participants had to selectively perform the age judgment on only one of the categories. On the first trial of each run, participants were explicitly told what category to begin judging (e.g., "start with faces"). The age (young or old) of the first trial defined the age of the current "mini-block" and participants were instructed to continue judging the same category as long as the age in that category stayed the same. Upon encountering a trial in which the age in the judged category was different (e.g., a change from "young" to "old"; see Figure 1A), the task rules required participants to switch to judging the age of the other category, starting a new mini-block on the next trial.

Performance of this task therefore required decision-making based on hidden states: in addition to (1) the observable current age, participants needed to know (2) which category they had to judge, (3) the age of the previous trial (except on the first trial of a mini-block), and (4) which category they had judged in the previous trial. The last aspect was necessary because changes in category signaled the beginning of a new mini-block, at which point it was not necessary to compare the age of the current trial to that of the previous trial (Figure 1C). The task was thus fully characterized using 16 states defined on these four binary-valued features (Figure 1D). Moreover, the mini-block structure meant that trials could be categorized as either the beginning

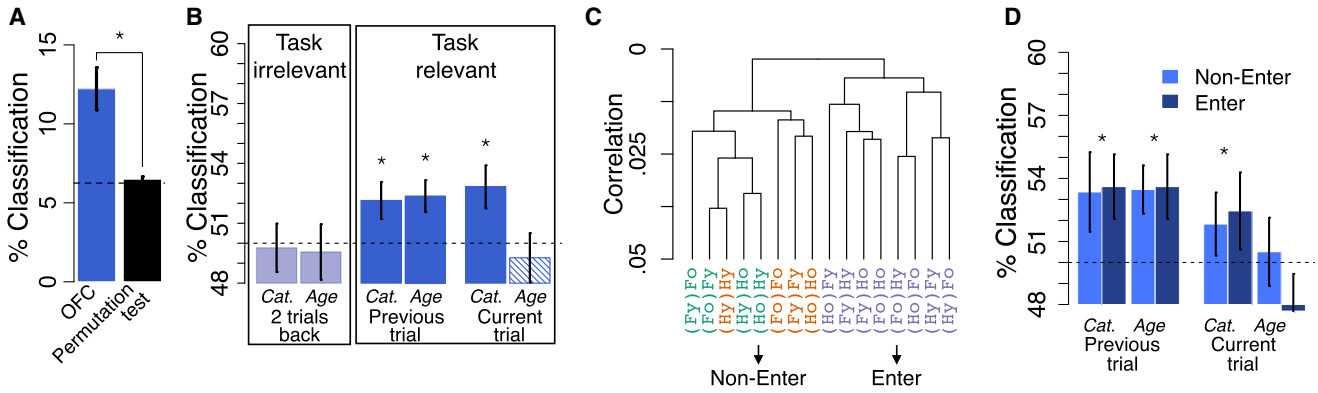

Figure 2. Encoding of Hidden State Information in OFC BOLD Signals

(A) Average 16-way classification of state identity from fMRI patterns within the anatomically defined OFC (blue bar) and following a permutation test (black bar).

(B) Contribution of information to 16-way classification separately for task-irrelevant information from two trials ago (leftmost bars, light blue), the three different hidden state components (solid blue), as well as the observable state component (striped bar). Only hidden and task-relevant components of the state contributed significantly to state identity decoding.

(C) Dendrogram indicating the similarity structure of different states according to a hierarchical cluster analysis. Colors and acronyms as in Figure 1D. Note that all purple (Enter) states involve a category switch, whereas all other states do not.

(D) Decoding of hidden state information separately for switch and non-switch trials. Dashed horizontal lines, chance baseline; error bars, SEM. *$p \leq 0.05$.

of a mini-block (Enter trials), within a mini-block (Internal trials), or the end of a mini-block (Exit trials, see Figure 1 and below). Therefore, although the display was highly similar in all trials, different mental states and mental operations were required in different trials as a consequence of our task rules.

Participants performed the task with near optimal performance (mean error rate: 2.3%; Figure S1), suggesting that they succeeded in mentally representing the correct state information. We therefore reasoned that if OFC encodes the current state of the task, we should be able to decode this information from OFC activity.

Using multivariate linear support vector machine classification (Chang and Lin 2011) on activity in the anatomically defined human OFC (Tzourio-Mazoyer et al., 2002; Kahnt et al., 2012), we could decode significantly above chance in which of the 16 possible task-states participants were during the task (mean classification accuracy 12.2%, i.e., almost twice the 6.25% chance baseline, $t_{26} = 9.04$, $p < 0.001$, Figure 2A). Next, we tested the prediction that information encoding in OFC is specific to task-relevant state components by constructing an alternative 16-state space that included task-irrelevant information about the category and age from two trials ago, instead of the category and age from one trial back. Training and testing a classifier on this state space resulted in lower 16-way classification compared to the hypothesized state space (8.1% versus 12.2%, $t_{26} = 2.8$, $p = 0.01$). To further test what information was represented in OFC, we then assessed the accuracy of the two 16-way classifiers in predicting each component separately (see the Experimental Procedures; note that this did not involve training new classifiers, but rather accounting for which components were incorrectly versus correctly classified on each test trial). Figure 2B shows the resulting accuracies. This analysis confirmed that no information about irrelevant events from two trials ago could be detected (two leftmost bars in Figure 2B, $p > 0.66$). In contrast, the category and age from one trial ago, as well as the current category—all hidden and task-relevant components—were classified above chance, as we had predicted (middle solid blue bars in Figure 2B, $t_{26} > 3$, p values for previous category, previous age and current category were 0.014, 0.003, and 0.007, respectively; p values one-sided). Finally, the analysis revealed no evidence that the current age, which was task-relevant but observable given the current category, contributed to state representations in the OFC (rightmost bar in Figure 2B, $p > 0.72$, see also Figure S2B).

It is interesting to note that we did not find any evidence that the observable component "current age" could be classified, while its unobservable counterpart, "previous age," was decodable in OFC. This finding raises the possibility that participants encoded the task using eight states (each consisting only of previous category and age, plus current category) rather than 16. Under this scheme, the current age determines the behavioral response and the transition to the next state, but is not part of the state space itself. Decoding eight different states in the OFC in this manner, we indeed found significant decoding accuracy, albeit at a slightly lower level relative to chance baseline (16.7%, corresponding to 4.2% above chance baseline, $t_{26} = 3.3$, $p = 0.002$); all components involved in this reduced state space also showed above baseline decoding (53.5%, 52.9%, and 55.5% for previous category, previous age, and current category, respectively; Figure S2C). We note that although this was not our a priori hypothesized state space, in RL it is common to separate actions from states and to predicate state transitions both on the current state and the current action. To further probe whether the current action was encoded as part of the state space, we also explored alternative state spaces that included the response mapping (that was determined randomly on each trial and displayed on the screen throughout the trial) or the current motor action (that was orthogonal to the current age due to the random response mapping) instead of the current age. Results of these analyses suggested the response mapping and

motor action were coded primarily in visual and motor cortex, respectively (see the Supplemental Information and Figure S3).

In order to rule out contributions of biases in the analysis, task design, or behavior to our decoding results, we scrutinized our results with several control analyses. First, a permutation analysis showed at chance decoding levels and was significantly lower than decoding on the real data, both in the 16-way analysis as well as regarding the binary classification (see black bar in Figures 2A and S2A, all $p < 0.025$). In addition, we repeated our decoding analyses on synthetic fMRI data that reflected the subject-specific time course of events and included univariate stimulus-driven activation within OFC (synthetic data were matched as closely as possible to real data in all other respects, see the Experimental Procedures). This analysis also showed no above-chance decoding (mean accuracy across simulated "participants" 6.8%, $p = 0.43$ when compared to the chance baseline of 6.25%). Based on these control analyses, we can confidently interpret our classification results as reflecting encoding of task states in the OFC.

Overall, the results above indicate that within OFC, different neural activation patterns were related to the identity of different hidden states. Such a detailed representation of the state space does not preclude the possibility that higher-level distinctions are additionally encoded in the neural signal. For instance, it would be possible that the neural codes for all Enter states (that involve a mental category switch) are more similar to each other than to non-Enter states (that do not involve such a switch). At the same time, individual Enter states could still be distinguishable from each other, reflecting a hierarchically nested state space representation. In order to explore whether such a hierarchical information structure exists, we investigated the correlations between multivoxel activity patterns associated with each pair of states (correlations were within subjects and between runs to avoid confounds of temporal proximity of states). The resulting similarity matrix (1 minus correlation, see Kriegeskorte et al. 2008) was then submitted to a hierarchical cluster analysis (see Farovik et al. 2015 and McKenzie et al. 2014 for a similar approach). The dendrogram summarizing the results (Figure 2C) indicates that pairs of activity patterns were indeed more similar if two states were both Enter or both non-Enter states ($p < 0.01$; paired t test comparing each subjects' average within-class versus between-class correlations). This structure is fully compatible with and complementary to the encoding of a detailed state space in which full hidden-state identity is represented, as confirmed by a decoding analysis done separately for Enter and non-Enter trials that showed that even within these subsets of trials the 16-way classifier could successfully detect all three hidden state components (Figure 2D, all $t_{26} > 2$, $p < 0.05$, note that only test sets, not training sets, were changed for this classification analyses, see the Experimental Procedures).

Next, we assessed the anatomical specificity of hidden state encoding by performing 16-way searchlight classification analyses across the whole brain (Kriegeskorte et al., 2006) and then evaluating their accuracy on each of the four state components (all analyses followed the same procedures used for the anatomical OFC analysis, see above). As can be seen from the resulting statistical brain maps (Figures 3A–3D; Table S1), a cluster in medial OFC was the only brain area in which decoding for all three hidden state components could be found. This contrasts, for example, with the anatomical diversity of encoding age and decoding category in more posterior sensory areas (see clusters in lingual gyrus for past as well as previous age and category encoding in parahippocampal "place area" and fusiform "face area"). Crucially, however, a voxel-wise conjunction analysis on the information maps of hidden state components confirmed that, across participants, only medial OFC activity carried complete information about all three hidden state components (Figure 3E; $p_{conj} < 0.01$, uncorrected, minimum cluster size 5 voxels, peak in left medial frontal orbital gyrus, Brodmann area 11, Montreal Neurological Institute [MNI] coordinates [3/44/−14], $Z_{peak} = 3.14$; see Table S1 for full results and Figure S3 for further analyses of the spatial distribution of information encoding in individual participants and signal-to-noise ratio maps within OFC). The uniqueness of the pattern of information encoded in OFC was also evident when compared to putatively task-relevant areas such as hippocampus (episodic memory), dorsolateral PFC (DLPFC; working memory), or fusiform face area and parahippocampal place area (FFA/PPA; representation of faces and houses, respectively), which all coded for current and previous category information but not previous age. None of these areas showed either complete encoding of all hidden state components nor specific encoding of only task-relevant variables (see Figures S2D–S2F).

Representing the correct state is an important prerequisite for successful decision-making. We therefore assessed the relationship between decoding accuracy and task performance across participants. Although behavioral errors were relatively rare due to pre-training and performance bonuses (see the Experimental Procedures), participants' average error rates were significantly correlated with the mean classification accuracy of all three hidden state components, suggesting a link between OFC state representations and task performance (Pearson correlation $r = -0.58$, $t_{25} = -3.6$, $p = 0.002$, Figure 4A; bootstrapped 95% confidence intervals [CIs]: [−0.8432, −0.0169], see Figure S4, same correlation with decoding in hippocampus or DLPFC not significant, $p > 0.14$, but a significant relation was found for PPA/FFA, $r = -0.40$, $t_{25} = -2.2$, $p = 0.03$).

Additionally, we investigated classification of task states on trials in which behavioral errors occurred. Testing the trained classifier on single-trial fMRI data, we found that incorrect responses were accompanied by worse classification—the average 16-way classification in the five trials preceding an error was 11% whereas accuracy was only 4.2% on error trials ($t_{23} = 2.2$, $p = 0.04$, for a comparison of the two; three participants excluded due to the lack of behavioral errors or for technical reasons, see the Experimental Procedures). The trial-by-trial classification time course revealed that decoding accuracy decreased in the trials preceding behavioral errors, that is, there was a negative linear effect of trial on decoding in the trials leading up to and including an error, as assessed by a mixed-effects model in which the classifier accuracy was the dependent variable and a linear effect of trial number relative to the error was tested ($\chi^2_1 = 4.7$, $p = 0.03$). No such decrease was detected in matched decoding time-courses that did not include errors (interaction of condition by trial, $\chi^2_1 = 4.3$, $p = 0.04$; effect of trial in non-error decoding, $p = 0.33$, see Figure 4B, filled circles). This effect of

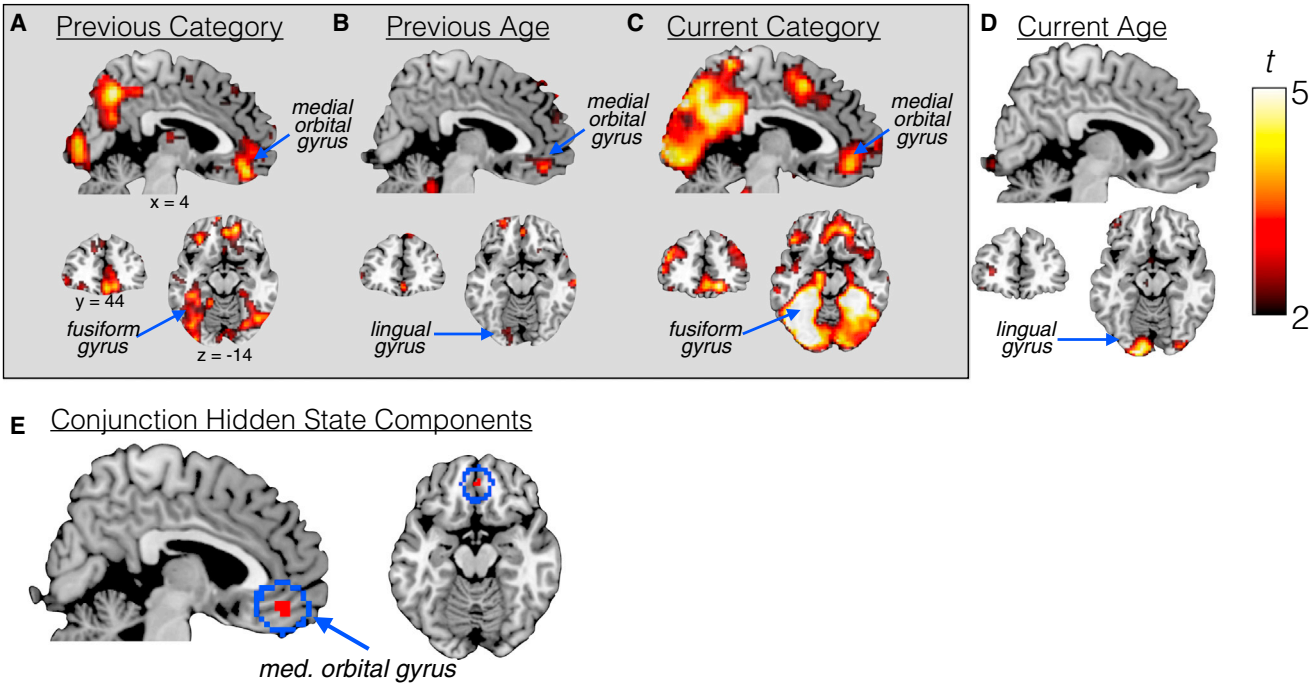

Figure 3. Anatomical Specificity of Hidden State Representations

(A–D) Searchlight maps for decoding accuracy (t test against chance baseline) for each of the four state components (A) previous category, (B) previous age, (C) current category, and (D) current age, thresholded at p = 0.025 (uncorrected) for illustration purposes. Colors represent t values. Sagittal, coronal, and axial slices are at x = 4, y = 44, and z = −14 (MNI), respectively.

(E) Conjunction analysis of three hidden components, showing that all three hidden components could be decoded simultaneously only within medial OFC (conjunction threshold $p < 0.01$, i.e., every state component is significant with $p \leq 0.01$, uncorrected, at shown location). Red voxels indicate significant searchlight centers; blue line indicates the outline of the corresponding searchlights. A complete table of results can be found in Table S1.

behavioral errors was also reflected in the fact that decoding in the error trial was significantly lower than decoding in the matched correct trials (trial "0" in the figure; $t_{24} = 2.3$, p = 0.01) and marginally lower in the trial prior to the error (trial "−1" in the figure; $t_{23} = 1.5$, p = 0.07). These findings of a transient drop in classification around the time of behavioral errors suggest that a gradual loss of state information in OFC foreshadowed performance errors in the task.

If task-state encoding in OFC is important for task performance, one might expect transitions between states whose representations are very different to be more error-prone than transitions between neurally similar states. We therefore investigated whether similarity relations among OFC activity patterns are correlated with task performance. We computed Pearson correlations between multi-voxel activity patterns associated with each pair of states (Figure 5A; correlations are within subjects and between runs to avoid confounds of temporal proximity of states; only positive correlations were considered because negative correlations were rare and difficult to interpret as a measure of distance). We then correlated these "distances" (1 minus the correlation) with performance accuracy, such that the neural similarity of a pair of consecutive trials A → B was correlated with the error rate on trial B. This was done for a homogenous subset of transitions—trials spanning a category switch following a change in age (i.e., the last trial of one mini-block and the first trial of the next mini-block)—to allow a fair comparison of error rates. Results showed that transitions to "nearer" states (i.e., transitions between pairs of states whose neural representations were more highly correlated) were associated with lower error rates (linear effect of correlation on error rates in the eight possible transition types: $t_6 = -3.23$, p = 0.02, within-subject mixed effects analysis: $\chi^2_1 = 3.0$, p = 0.08, Figures 5B and 5C). No such relationships were seen in data from hippocampus, DLPFC or FFA/PPA ($p > 0.17$ for all). Finally, as with our decoding analyses, we tested the validity of our results by performing the same analyses on synthetic fMRI data that included univariate stimulus-driven activation within OFC and were matched closely to real data in all other respects. These control analyses found no significant correlations with performance ($p > 0.18$ for all), allowing us to conclude that our results cannot be explained by temporal contingencies or univariate stimulus-driven activation. In addition, we repeated this same analysis using Euclidean distances instead of correlations and obtained the same results (see Figure S5), showing the robustness of our finding with respect to different distance measures. Hence, our results support the idea that the OFC contains a multivariate encoding of state identity that is used for task performance.

DISCUSSION

Taken together, our results demonstrate that patterns of fMRI activity in human OFC contain information about participants'

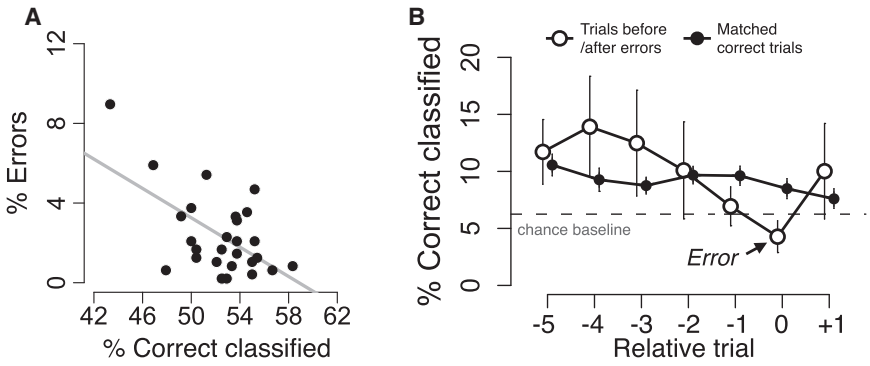

Figure 4. OFC State Classification Correlates with Task Performance

(A) A significant relationship between average classification accuracy within OFC (mean across all three hidden state components) and participants' error rate suggests that the encoded state information was relevant for task performance. Each dot represents one participant. Gray, regression line.

(B) Trialwise decoding before, during, and after behavioral errors. Empty circles, 16-way decoding accuracy in the five trials leading up to, during and after an error. Filled circles, decoding accuracy during seven consecutive trials with no behavioral error. Chance = 6.25%, error bars, SEM.

current location in a mental map of a task comprised of hidden states. Multivariate classification of human brain activity during a 16-state decision-making task showed that unobservable information about the task environment could be reliably decoded from OFC activity patterns. Moreover, only task-relevant state components were represented in OFC and the fidelity of state encoding, as well as the similarity between representations of different task states, were related to participants' task performance. These results support the proposal that OFC encodes hidden states in the service of RL.

Interestingly, state information encoded in OFC activation patterns did not include "current age," despite it being a salient aspect of the task that was needed to make the correct response. One distinguishing feature of current age was that it was perceptually visible, whereas the other state components of previous and current category and previous age were not. Another difference is that current age may be considered more of an action rather than a state component—it determines the current response and the transition to the next state, but does not necessarily have to be encoded as part of the state itself. Indeed, this same non-decodable current age was decodable in the OFC one trial later, when it was the previous age. Based on the results of studies in which the OFC was compromised, we had previously suggested that the OFC is especially critical for representing hidden states (that is, disambiguating observable states based on unobservable information) (Wilson et al., 2014; Bradfield et al., 2015). However, it is not immediately clear how this hypothesis at the level of states should be translated to hidden or observable state components as there are several possible implementations for an OFC-based signal that disambiguates otherwise similar states. Moreover, observable states may also be represented in the OFC, but since OFC lesions did not impair tasks that relied only on such states, these are probably also represented elsewhere in the brain. The question of the precise implementation of the state representation in OFC, and how it is combined with state information in the striatum, is therefore an interesting avenue for future work.

Our proposal that the OFC represents a cognitive map of task-state space offers an integrating framework for existing OFC theories, as different tasks can require that states include information about emotional valence (Bechara et al., 2000), reward and punishments (Kringelbach, 2005) and state expectancies (Schoenbaum et al., 2011). For example, the plethora of evidence showing value signals in OFC (Padoa-Schioppa and Assad, 2006, 2008) could be due to the fact that expected reward often represents an important and hidden aspect of the task and therefore is part of the task state. Similarly, OFC's role in delayed match-to-sample working-memory tasks (Meunier et al., 1997) and n-back tasks (Barbey et al., 2011) may be attributable to the fact that decision-making in these tasks relies on hidden state (working memory) information. This may also explain previous fMRI findings that showed a general increase in OFC activity in working memory or recognition memory tasks (Lamar et al., 2004; Schon et al., 2008; Frey and Petrides 2000). In addition, studies showing that value-related signals in OFC are modulated by context (Winston et al., 2014) are in line with our idea that state representations integrate value signals with other task-relevant information. Despite the potential breadth of this framework, it is important to note that our view emphasizes OFC's role in representing task states in decision-making and reinforcement-learning tasks in particular. This implies that we would not necessarily expect OFC to represent working memory in all tasks, but rather only in tasks that require decision-making and learning based on working memory. Future studies are needed to evaluate the full scope of OFC's role. In this regard, lack of involvement of the OFC in previous studies may not provide conclusive evidence against a role for OFC in the explored tasks, as an involvement of OFC in representing the states of a task would not necessarily be reflected in overall higher activity. Moreover, due to susceptibility and dropout artifacts, previous studies using protocols that were not specifically geared at acquiring OFC signals may lack data that can reliably adjudicate regarding multivariate encoding of states in OFC.

While task states may seem a permissively flexible construct, our hypothesis suggests that for any given task these representations will not be promiscuous—because RL mechanisms must learn values for each action at each state of the environment, the efficiency of learning and decision-making scales badly with an increasing number of states. As a result, a task should be represented with a small but sufficient set of states, implying that only task-relevant state information should be encoded in the OFC, at least to the extent that the animal has enough familiarity with the task to know the correct state representation (Gershman and Niv, 2010; Niv et al., 2015). Indeed, in our fully instructed task, OFC selectively represented task-relevant information (Figure 2A), possibly curated from input from its wide network of

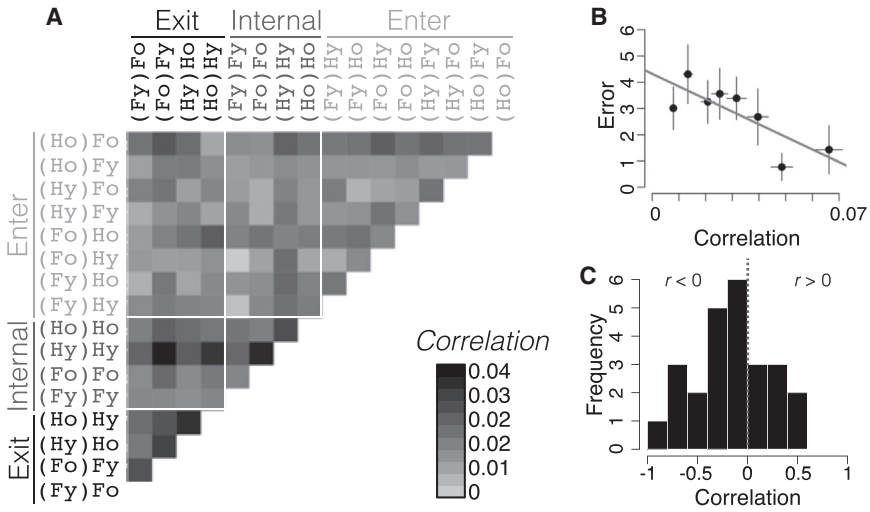

Figure 5. OFC State Representations Affect Task Performance

(A) Average correlations between neural state representations within OFC. Darker gray denotes higher correlation (i.e., more similar state representations).

(B) Relationship between error rate on the eight different transitions exiting one miniblock and entering another and correlation between the pairs of states corresponding to these transitions, across participants. For each participant, the eight transitions were ordered according to strength of correlation (from low to high). Dots denote the average correlation between states in that ordinal position across participants (x axis) and average behavioral error rate on the corresponding transitions (y axis), with horizontal and vertical error bars denoting SEM of each. Higher correlations between neural states were associated with fewer behavioral errors, on average ($p = 0.04$).

(C) Histogram of within-subject correlations between error rates and neural state similarity showing that correlations were, on average, significantly lower than 0 ($p < 0.01$).

connected brain areas that represent candidate state components more broadly. Recent studies demonstrating that OFC encodes reward identity (Howard et al., 2015), but generalizes over task-irrelevant stimuli (Klein-Flügge et al., 2013), are in line with this notion that orbitofrontal task states are influenced by task relevance (Stalnaker et al., 2015). Importantly, all these accounts emphasize OFC's role in decision-making, and our proposal that OFC may, in principle, represent any of these types of information (in a task-dependent manner) therefore subsumes several seemingly conflicting hypotheses and integrates existing findings.

Another interesting question for future research regards the role of value representations in OFC, which might be reflected in hierarchical representations that integrate value and other state information (Farovik et al., 2015). Indeed, the precise link between state information in OFC and value computations in the basal ganglia has yet to be delineated. In addition, our findings sidestep questions about the anatomical organization within OFC and do not preclude the idea that this large and physiologically heterogeneous area (Ongür and Price, 2000) contains multiple, functionally different networks (Kringelbach and Rolls, 2004). In particular, the fact that across participants, state representations localized in medial OFC should be interpreted with caution as medial OFC enjoys relatively less signal dropout as compared to lateral OFC and a larger concentration of gray matter within a spherical searchlight. The large between-subject variability that we observed (Figure S3) and the known inter-subject variability in sulcal patterns (Kringelbach and Rolls, 2004) also call for caution in this regard.

Moreover, we do not propose that state information be exclusive to the OFC. Because task states are necessary for correct performance, presumably state representations in the OFC would be conveyed to downstream areas for both model-based and model-free decision-making (Daw et al., 2005). In addition, whereas we suggest (and present data supporting) a role for the OFC in providing a carefully curated summary of state-related information, this representation no doubt reflects information available elsewhere in the brain. Indeed, the OFC's wide-ranging connectivity to sensory, associative, and high-order areas can readily support such a function. Therefore, decoding task-relevant information in other brain areas is not antithetical to our proposed role for the OFC. Indeed, other brain areas have been shown to dynamically and selectively represent task-relevant aspects during decision-making (Schuck et al., 2015)—the interaction between OFC and these frontal areas is at present still unclear.

Finally, the demonstration that state representations that are divorced from immediate sensory input and are important for decision-making can be decoded from noninvasive brain-imaging data opens the door to future work that will provide a read-out of such representation, and use it to give corrective feedback to wrong representations and not only to wrong actions. This can allow for better training and teaching and will more directly establish the causal relationship between OFC representations and decision-making behavior. Our previous work has already shown that using decoding techniques to assess what information about the environment is selectively represented allows insight and prediction of qualitative differences in how humans approach decisions and which aspects of the environment they will learn about (Schuck et al., 2015). Because the specific way in which a task has been mentally encoded is often difficult to assess from behavior, but is critical for the success of a decision maker, a direct measurement of these representations as shown here holds immense promise both for basic science and for practical applications.

EXPERIMENTAL PROCEDURES

Participants

Thirty adults from the Princeton University community participated in exchange for monetary compensation ($20 per hour plus up to $10 performance-related bonus). Two participants were excluded from further analysis due to changed

fMRI settings and one participant was excluded due to a large number of errors (>3 SDs of group mean). The remaining 27 participants (13 female) had a mean age of 22.18 years (range 18–31), were right-handed, had normal or corrected-to-normal vision, no psychiatric illnesses, and fulfilled all standard eligibility criteria for participation in an fMRI study. The study was approved by the Princeton Institutional Review Board and all subjects gave informed written consent prior to participation. Because behavioral data indicated that learning of the task rule had reached asymptote by the time of scanning (i.e., participants did not show signs of changes in error rates or reaction times across time, see Supplemental Information), we included participants even if they had not completed all five runs (three participants had fewer runs due to timing restrictions or technical errors; mean number of runs completed: 4.9). For the analysis of trialwise decoding around errors, we had to exclude three additional participants: two because time stamps for behavioral errors were lost due to an error in the code that caused these trials to be overwritten and one whose errors only occurred at the beginning of scans, such that the decoding time course of five trials preceding the behavioral error could not be reconstructed.

Stimuli

Each stimulus consisted of spatially superimposed images of a face and a house (images courtesy of Dr. Dorothea Haemmerer and http://faces.mpdl.mpg.de/faces (Ebner et al. 2010) (Figure 1A). Faces and houses could be classified as either young or old, i.e., face images showed either a younger or older adult, and house images showed either a contemporary (i.e., young) or old-fashioned (i.e., old) building. This resulted in four possible classes of stimuli, two combining age-congruent pictures (both face and house are either young or old) and two combining age-incongruent pictures (an old face combined with a young house or vice versa).

Task

The task was structured into mini-blocks, which required participants to judge the age of either faces or houses. Within each mini-block, participants' task was to judge the age of a given category, while switching the judged category between mini-blocks. To start the task, a cue indicated the category to be judged on the first mini-block (e.g., faces). Then, the mini-block continued with judging the same category as long as the age of the object in this category did not change. Once the age changed (from young to old or vice versa), participants were required to start judging the other category (e.g., houses instead of faces) from the following trial, effectively starting a new mini-block. Thus, face and house mini-blocks alternated (see Figure 1A). Importantly, category switches were not cued. Each mini-block lasted at least two trials (average block length: three trials, 32 mini-blocks/switches per scanner run). No age comparison was required between the first trial of the new mini-block and the last trial of the previous mini-block. These rules created three basic types of trials: Enter trials, which started a new mini-block, Internal (within-mini-block) trials, in which the age repeated and subjects had to continue judging the same category, and Exit trials, which had a different age than the previous trial and thus signaled the need for a category switch (and a new miniblock) on the next trial (Figure 1C).

At the start of each scanner run, a cue was displayed indicating the first category to judge (4 s). Then each trial began with the display of the young/old response mapping below a fixation cross (mean duration: 1.2 s, range: 0.5–3.5 s), after which the overlaid face-house stimulus was displayed (mean duration: 3.3 s, range: 2.75–5 s; Figure 1B). This resulted in an average trial duration of 4.5 s (range: 3.25–8.5 s; all intervals drawn from a truncated exponential distribution). Participants responded "old" or "young" by pressing the left or right button with the index or middle finger of their right hand. The mapping between the response keys and old or young age changed trialwise, ensuring an equal number of left/right presses for both ages. Participants had up to 2.75 s to indicate their response. The chosen age was indicated by a small rectangle, however, the stimulus stayed on the screen until the end of the stimulus presentation duration, such that stimulus and trial durations were independent of reaction times. Erroneous or time-out responses led to feedback (0.7 s) and, for Internal and Enter trials, a repetition of the trial with the correct object category displayed on the screen. Because error-triggered repetition of Exit trials would require a category switch immediately after the repeated trial, the trial preceding the error was repeated in these cases (also with the correct category displayed on the screen), avoiding the need for a category switch immediately following the repetition. No feedback was given otherwise, but at the start of the scanning session participants were told that they would receive 1¢ per correct answer plus a $5 bonus if their overall error rate stayed below 2.5%.

Design

Participants were first trained on the task outside the scanner (three blocks, 97 trials each, same week as main task). On the day of scanning, participants first practiced a simplified version of the task with non-overlapping face/house images while lying in the scanner. The main task was then started, which consisted of five blocks of trials (485 total), performed during fMRI acquisition. Sequences of stimuli were selected such that the four image types (old or young faces/houses), appeared equally often on the attended and unattended dimension in each block. The age of the face and house was congruent in one-third of all trials. Within each scanner run, only two specific images for each age and category were used to construct all stimuli. Each block consistent of 32 same-category-same-age, 32 same-category-different-age trials, 16 different-category-same-age, and 16 different-category-different-age trials. Enter, Internal, and Exit trials each had an equal number of young/old faces or houses, and in any trial there was a 50% chance that the age changed on the next trial. The less crucial transitions on the non-attended dimension were matched approximately, with a maximum deviation from equal probabilities by 5% (e.g., 55/45 instead of 50/50).

State Space and Markov Property

As we report in the main manuscript, successful performance of the task required participants to maintain four pieces of information—the current and previous ages and categories. Each combination of these four aspects of the task environment constitutes a state, which we denote by its acronym: a state labeled (*Ho*)*Fy* indicates a trial in which the previously judged category was house (*H*), the previous age was old (*o*), the current category is face (*F*), and the age of this face is young (*y*). Because each of the four components of a state could take one of two values (ages could be *o* or *y* and categories *H* or *F*), this resulted in a total of $2^4 = 16$ states. The rules we imposed on our task implied that each state could transition to exactly two other states with equal transition probability (32 possible transitions). Importantly, only one state component was observable from the current visual input (the current age), whereas all others were hidden and depended on memory of the previous trial and knowledge of the task rules. The so-defined states form a Markov decision process as states are only dependent on their immediate predecessor. Formally, the Markov property is fulfilled if the joint probability of the next state (s_{t+1}) and next outcome (r_{t+1}) depends only on the current state (s_t) and action (a_t) and is independent of all preceding events:

$$p(s_{t+1}=s, r_{t+1}=r \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = p(s_{t+1}=s, r_{t+1}=r \mid s_t, a_t). \quad \text{(Equation 1)}$$

Scanning Protocol

MRIs were acquired at the Princeton Neuroscience Institute using a 3-T Siemens Magnetom Skyra MRI scanner (Siemens) and optimized for imaging the orbitofrontal cortex (Weiskopf et al., 2007). A T2*-weighted echo-planar imaging (EPI) pulse sequence was used for functional imaging (3 × 3 mm in plane resolution, slice thickness = 2 mm, gap = 50%, TR = 2,400 ms, TE = 27 ms, FOV = 196 mm, flip angle = 71°, 46 axial slices, interleaved acquisition, 64 × 64 matrix). Slice orientation was tilted 30° backward relative to the anterior-posterior commissure axis. After the experiment, field maps for distortion correction were acquired using the same parameters (TE1 = 3.99 ms) and structural images were acquired with a high-resolution MPRAGE pulse sequence (voxel size = 0.9 × 0.9 × 0.9 mm). Participants' respiration and pulse was acquired during scanning using pulse oximetry and a pneumatic respiratory belt. The experiment began 11 s after acquisition of the first volume of each run. The temporal signal-to-noise ratio (voxelwise mean ÷ voxelwise SD) is shown in Figure S3D.

Data Preprocessing

Standard fMRI data preprocessing and analyses were done using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and the PhysIO Toolbox for SPM (http://www.

translationalneuromodeling.org/tapas/). Multivariate pattern analyses were done using the LIBSVM implementation of support vector machines (Chang and Lin, 2011). Software is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm and the Princeton MVPA toolbox (http://code.google.com/p/princeton-mvpa-toolbox). Behavioral analyses and computations within the assumed graphical model of state space (see below) were done using R (R Core Development Team, 2014).

The first trial within each scanner run, error trials, and trials following errors were excluded from all analyses other than error-rate analyses. fMRI data preprocessing followed standard procedures and involved spatial realignment, coregistration of anatomical and functional scans, segmentation of structural scans into white and gray matter maps, normalization into MNI space, and smoothing. First-level (subject-wise) general linear modeling involved regressors of interest that captured stimulus onset events (see below) and nuisance regressors that reflected participant movement (six regressors) as well as cardiac phase (following Glover et al., 2000, as implemented in the PhysIO Toolbox; 26 regressors). Trial events were modeled as boxcar functions whose length reflected the reaction time in that particular trial to account for reaction time-induced variability (Grinband et al., 2008). All regressors were convolved with a canonical hemodynamic response function. Anatomical regions of interest (ROIs) were created using SPM's wfupick toolbox. The OFC ROI was defined as in Kahnt et al. (2012) and included bilateral inferior, middle, and superior orbital gyri and bilateral rectal gyri according to the automated anatomical label (AAL) atlas. The hippocampus (HC) was defined as the left and right hippocampus AAL labels. Dorsolateral prefrontal cortex (DLPFC) was defined as the middle frontal gyrus AAL labels. The PPA/FFA mask included bilateral fusiform and parahippocampal gyri.

fMRI Data Analysis

Standard general linear models (GLMs) were used to estimate voxelwise activations associated with stimulus display. First-level models were run on realigned but non-normalized, non-smoothed data and included separate regressors for each of the 16 different states plus the above described 32 nuisance regressors for movement and physiological noise and the runwise mean activation. This resulted in five wholebrain maps of parameter estimates ("betas") for each of the 16 states (one map for each state and run). For classification analyses, we *Z* scored and spatially smoothed the resulting beta maps within each run and used them as labeled examples. We applied a leave-one-run-out cross-validation scheme to train and test a support vector machine (SVM, linear kernel, cost parameter = 1) on examples of the 16 different states. For the main analysis reported in Figure 2A, the test set comprised the run-wise beta estimates for each state. To obtain the results in Figure 2D, the test set was split into Enter and non-Enter trials, and the classification test was done separately for each subset. These separate classification results each characterize how well decoding can be done in the absence of Enter/non-Enter differences. Because each test was noisy due to the small number of test points, results were then averaged within participant to maintain statistical power. For the trial-wise decoding shown in Figure 4B, the test set consistent of raw data from the relevant trials, spatially smoothed (FWHM = 3 mm) and *Z* scored run-wise. For each trial, data recorded 2 TRs after stimulus onset (4.8–7.2 s) were used as input to the 16-way classifier, and decoding accuracy was measured. That is, the training set and the classifier were identical for the run- and trial-wise analyses, which differed only in their test set.

All classification analyses resulted in a predicted state of each test example that could then be compared against the true state. Apart from the basic test of whether the predicted state aligned with the true state (i.e., assessing 16-way classification accuracy), we decomposed these predictions into decoding accuracies of the four components of the state space by considering the agreement between predicted and true state on each component separately. For example, a classifier might wrongly predict class (*Fy*)*Fy* for an (*Ho*)*Fy* test item. Although this prediction is overall wrong, the current category and the current age (*Fy*) are predicted correctly, and only the components that relate to the previous trial are wrong ($F \neq H$ and $y \neq o$). In this manner, we obtained the correctness of each test example on each of the four components and determined classification accuracy separately for each component. This approach has the advantage that it allowed us to base our analysis on separate patterns associated with each of the 16 states while at the same time it yielded classification for the four components. In the main manuscript, we refer to this decomposition of the classifier performance as "predicting each component separately." We did not attempt to classify each of the four dimensions using a binary classifier because that would be inconsistent with our hypothesis of separate representations for each of the 16 states. For example, for a direct classification of the current category eight states would be given the same label, however, we have no a priori reason to assume that such a joint representation exists in OFC.

State Space Similarities

The similarities between neural state patterns were computed as the Pearson correlations between the estimated (beta) maps of activities associated with each state. Correlating patterns estimated from the same run can introduce biases in the correlation matrix through the effect of temporal contingencies on the estimated correlation (Cai et al., 2016; Diedrichsen et al., 2011). To prevent this, we calculated pattern similarity by cross-correlating patterns estimated from separate runs, ensuring that the events modeled by each pair of regressors are fully temporally separated and therefore the regressors are orthogonal. Because the noise from two different runs is unlikely to be correlated, this method does not introduce biases into the estimated correlations of patterns (Alink et al., 2015). Specifically, for each run, each state map (masked by the anatomical OFC) was correlated with all state maps from all other runs. This resulted in four 16 × 16 correlation matrices, which were then averaged. This procedure was repeated for all five runs, and eventually the average correlation matrices for all five runs were averaged again. Negative correlations were set to NaNs as they are difficult to interpret and likely reflect noise in such a setting. Control analyses using synthetic data as well as Euclidean distances followed the same procedures described above.

Synthetic fMRI Data and Noise Simulations

To validate our results and analysis pipelines, we created and analyzed synthetic fMRI data. Specifically, we tested if univariate stimulus-related activation in OFC could explain our decoding or RSA results. Synthetic fMRI data were created in R using the package neuRosim, using the following specifications: for each subject and run, wholebrain activation was created. A single spherical activation peak in OFC was simulated (size drawn from normal distribution with mean of five voxels, SD 1, and fading parameter = 0.01). This peak was assumed to be activated by the presentation of each stimulus (using the actual onsets and durations for each subject). The location of each peak was set to be the location of the maximum decoding for each participant individually (see Figure S5) plus some spatial noise (mean = 3, SD = 2 voxels). Activation sizes for the 16 different states were drawn from a normal distribution (mean = 25, SD = 5) and thus were different for different states. Activation was convolved with a canonical HRF (double-gamma) function. In addition to the stimulus-related activation, whole-brain noise was simulated as a mixture of Rician system noise, temporal noise of order 1, low-frequency drift, physiological noise, and task-related noise using default mixture parameters and had a baseline value of 10. Spatial noise was modeled using a Gaussian random field. Signal-to-noise ratio was set to the default value of 2.87.

Following the simulation, the spatial smoothness of the corresponding real data were estimated using AFNI's 3dFWHMx function, and the synthetic 4D activation data were then smoothed to have the same spatial smoothness (using AFNI's 3dBlurToFWHM function). Finally, these data were preprocessed and analyzed in exactly the same manner as the real data.

Behavioral Analyses

Reaction times (RTs) reflect the median within each factor cell. Behavioral analyses were done using t tests or mixed-effects models using the package lmer in R. In the latter, participants were considered a random effect on the intercept and linear effects of trial or errors. The reported p values correspond to Wald chi-square (χ^2) tests. The analyses of the relationship between neural state similarity and error rates were also done using mixed-effects models. The significance of the correlation between average decoding and error rates shown in Figure 4 was tested using a bootstrapping approach (10,000 iterations, done in R using the package "boot" with default settings, see also Figure S4).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.08.019.

AUTHOR CONTRIBUTIONS

N.W.S., Y.N., and R.C.W. designed the research. N.W.S. conducted the research. N.W.S., Y.N., and M.C. analyzed data. All authors contributed to interpreting results and writing the manuscript.

ACKNOWLEDGMENTS

This publication was made possible through the support of NIH grant R01MH098861, Army Research Office Award W911NF-14-1-0101, funding from the Intel Corporation, and a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation, Army Research Office, or the NIH. We thank P. Dayan, A. Langdon, M. Sharpe, S. Chan, A. Radulescu, A. Geana, Y.S. Shin, and N. Rouhani for helpful comments on previous versions of the manuscript, and S. Chan and K. Holmes for help with data acquisition.

Received: March 27, 2016
Revised: July 11, 2016
Accepted: August 8, 2016
Published: September 21, 2016

REFERENCES

Alink, A., Walther, A., Krugliak, A., van den Bosch, J.J.F., and Kriegeskorte, N. (2015). Mind the drift - improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. bioRxiv, 032391.

Barbey, A.K., Koenigs, M., and Grafman, J. (2011). Orbitofrontal contributions to human working memory. Cereb. Cortex *21*, 789–795.

Bartra, O., McGuire, J.T., and Kable, J.W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage *76*, 412–427.

Bechara, A., Damasio, H., and Damasio, A.R. (2000). Emotion, decision making and the orbitofrontal cortex. Cereb. Cortex *10*, 295–307.

Bradfield, L.A., Dezfouli, A., van Holstein, M., Chieng, B., and Balleine, B.W. (2015). Medial orbitofrontal cortex mediates outcome retrieval in partially observable task situations. Neuron *88*, 1268–1280.

Cai, M.B., Schuck, N.W., Pillow, J., and Niv, Y. (2016). A Bayesian method for reducing bias in neural representational similarity analysis. bioRxiv. http://dx.doi.org/10.1101/073932.

Cavada, C., Compañy, T., Tejedor, J., Cruz-Rizzolo, R.J., and Reinoso-Suárez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex. A review. Cereb. Cortex *10*, 220–242.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 1–27.

Chase, H.W., Kumar, P., Eickhoff, S.B., and Dombrovski, A.Y. (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. Cogn. Affect. Behav. Neurosci. *15*, 435–459.

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

Dayan, P. (1993). Improving generalization for temporal difference learning: the successor representation. Neural Comput. *5*, 613–624.

Diedrichsen, J., Ridgway, G.R., Friston, K.J., and Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: a pattern-component model. Neuroimage *55*, 1665–1678.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. Nat. Rev. Neurosci. *2*, 820–829.

Ebner, N.C., Riediger, M., and Lindenberger, U. (2010). FACES–a database of facial expressions in young, middle-aged, and older women and men: development and validation. Behav. Res. Methods *42*, 351–362.

Farovik, A., Place, R.J., McKenzie, S., Porter, B., Munro, C.E., and Eichenbaum, H. (2015). Orbitofrontal cortex encodes memories within value-based schemas and represents contexts that guide memory retrieval. J. Neurosci. *35*, 8333–8344.

Frey, S., and Petrides, M. (2000). Orbitofrontal cortex: A key prefrontal region for encoding information. Proc. Natl. Acad. Sci. USA *97*, 8723–8727.

Gershman, S.J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. *20*, 251–256.

Glover, G.H., Li, T.Q., and Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. *44*, 162–167.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., and Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. Neuroimage *43*, 509–520.

Howard, J.D., Gottfried, J.A., Tobler, P.N., and Kahnt, T. (2015). Identity-specific coding of future rewards in the human orbitofrontal cortex. Proc. Natl. Acad. Sci. USA *112*, 5195–5200.

Kahnt, T., Chang, L.J., Park, S.Q., Heinzle, J., and Haynes, J.D. (2012). Connectivity-based parcellation of the human orbitofrontal cortex. J. Neurosci. *32*, 6240–6250.

Klein-Flügge, M.C., Barron, H.C., Brodersen, K.H., Dolan, R.J., and Behrens, T.E. (2013). Segregated encoding of reward-identity and stimulus-reward associations in human orbitofrontal cortex. J. Neurosci. *33*, 3202–3211.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

Kringelbach, M.L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. Nat. Rev. Neurosci. *6*, 691–702.

Kringelbach, M.L., and Rolls, E.T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Prog. Neurobiol. *72*, 341–372.

Lamar, M., Yousem, D.M., and Resnick, S.M. (2004). Age differences in orbitofrontal activation: an fMRI investigation of delayed match and nonmatch to sample. Neuroimage *21*, 1368–1376.

McDannald, M.A., Esber, G.R., Wegener, M.A., Wied, H.M., Liu, T.L., Stalnaker, T.A., Jones, J.L., Trageser, J., and Schoenbaum, G. (2014). Orbitofrontal neurons acquire responses to 'valueless' Pavlovian cues during unblocking. eLife *3*, e02653.

McKenzie, S., Frank, A.J., Kinsky, N.R., Porter, B., Rivière, P.D., and Eichenbaum, H. (2014). Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. Neuron *83*, 202–215.

Meunier, M., Bachevalier, J., and Mishkin, M. (1997). Effects of orbital frontal and anterior cingulate lesions on object and spatial memory in rhesus monkeys. Neuropsychologia *35*, 999–1015.

Niv, Y. (2009). Reinforcement learning in the brain. J. Math. Psychol. *53*, 139–154.

Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., and Wilson, R.C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. J. Neurosci. *35*, 8145–8157.

Ongür, D., and Price, J.L. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. Cereb. Cortex *10*, 206–219.

Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature *441*, 223–226.

Padoa-Schioppa, C., and Assad, J.A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat. Neurosci. *11*, 95–102.

R Core Development Team (2014). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing). https://www.R-project.org/.

Schoenbaum, G., Takahashi, Y., Liu, T.L., and McDannald, M.A. (2011). Does the orbitofrontal cortex signal value? Ann. N Y Acad. Sci. *1239*, 87–99.

Schon, K., Tinaz, S., Somers, D.C., and Stern, C.E. (2008). Delayed match to object or place: an event-related fMRI study of short-term stimulus maintenance and the role of stimulus pre-exposure. Neuroimage *39*, 857–872.

Schuck, N.W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P.A., Haynes, J.D., and Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. Neuron *86*, 331–340.

Stalnaker, T.A., Cooch, N.K., and Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. Nat. Neurosci. *18*, 620–627.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (Cambridge: Cambridge University Press).

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage *15*, 273–289.

Weiskopf, N., Hutton, C., Josephs, O., Turner, R., and Deichmann, R. (2007). Optimized EPI for fMRI studies of the orbitofrontal cortex: compensation of susceptibility-induced gradients in the readout direction. MAGMA *20*, 39–49.

Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. Neuron *81*, 267–279.

Winston, J.S., Vlaev, I., Seymour, B., Chater, N., and Dolan, R.J. (2014). Relative valuation of pain in human orbitofrontal cortex. J. Neurosci. *34*, 14526–14535.

Neuron, Volume *91*

Supplemental Information

Human Orbitofrontal Cortex Represents a Cognitive Map of State Space

Nicolas W. Schuck, Ming Bo Cai, Robert C. Wilson, and Yael Niv

Supplemental Information: Human Orbitofrontal Cortex Represents a Cognitive Map of State Space

Nicolas W. Schuck[1,*], Ming Bo Cai[1], Robert C. Wilson[2] & Yael Niv[1]

[1]Princeton Neuroscience Institute and Department of Psychology
Princeton University, Washington Road, Princeton, NJ, 08544, USA
[2]Department of Psychology
University of Arizona, 1503 E University Blvd, Tucson, AZ 85721

*Corresponding author contact:
Princeton Neuroscience Institute
Princeton University
Princeton, NJ, 08544, USA
email: nschuck@princeton.edu
tel: +1 609 258 7498

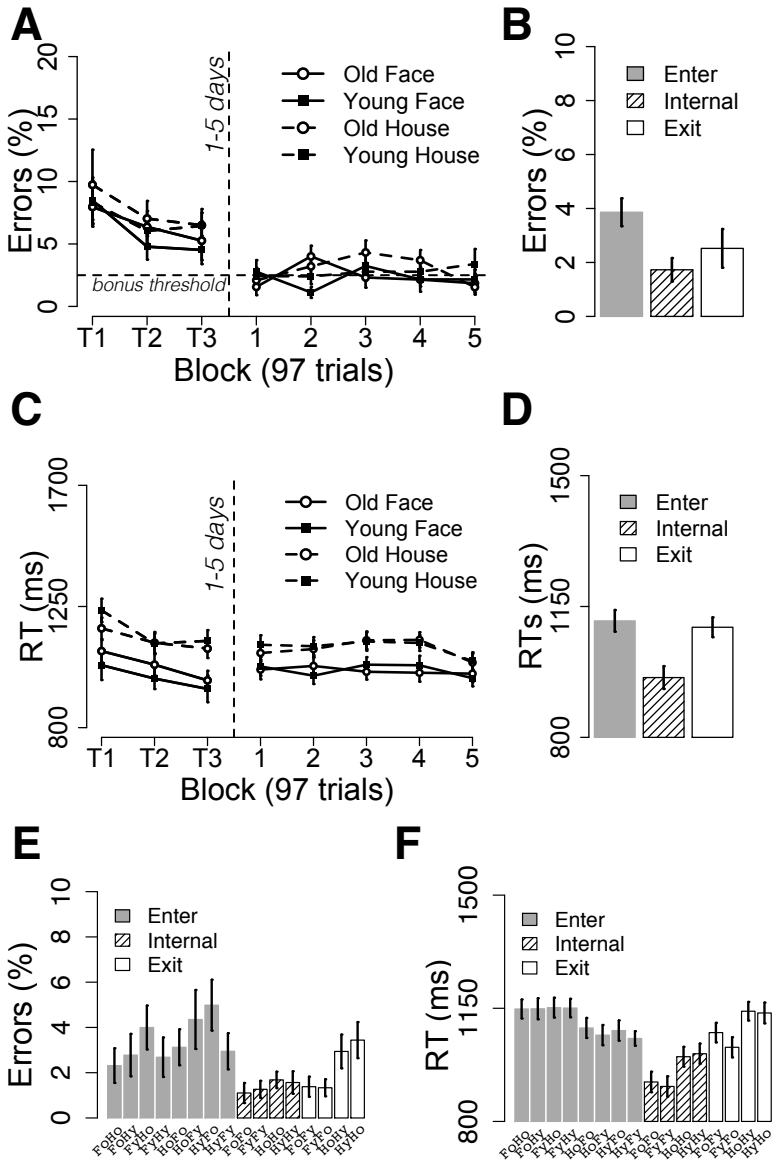

Figure S1. Behavioral results, related to Figure 1. *(A)*: Average behavioral error rate during training and the main experiment. The separate lines distinguish between categories (face = solid lines, house = dashed lines) and the age (filled circles = old, empty circles = young). Dashed horizontal line indicates the error level below which participants received a cash bonus in the scanning session. *(B)*: Average errors during the main experiment separately for Enter, Internal and Exit States. *(C+D)*: Average RTs during training and the main experiment and separately for Enter, Internal and Exit states, format as in (A) and (B), respectively. *(E+F)*: Average error rates and RTs for each of the 16 states during the main experiment. Error bars = S.E.M.

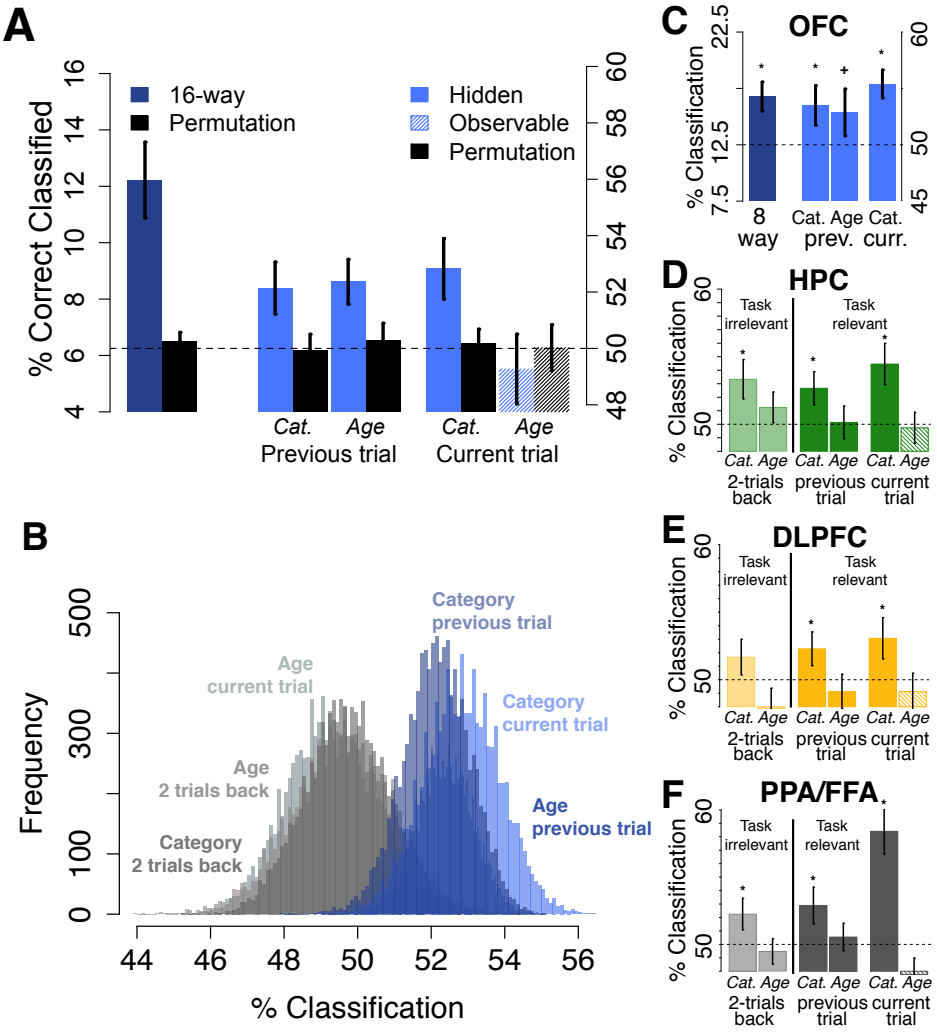

Figure S2. Decoding permutation test and 8-way decoding in different ROIs, related to Figure 2. *(A)*: Results of a permutation test (black bars) show greatly decreased decoding relative to the original decoding shown in Figure 2 (blue bars). *(B)*: Bootstrapped distributions of binary decoding for the six results shown in Figure 2B. *(C):* Eight-way decoding for which states with different current ages were modeled as the same event type in an anatomical ROI of OFC. *(D-F):* Component-wise decoding in different ROIs for comparison (format as in Figure 2B). The dashed horizontal line represents chance baseline, error bars represent ± SEM, *: $p < .05$, against baseline, one-tailed.

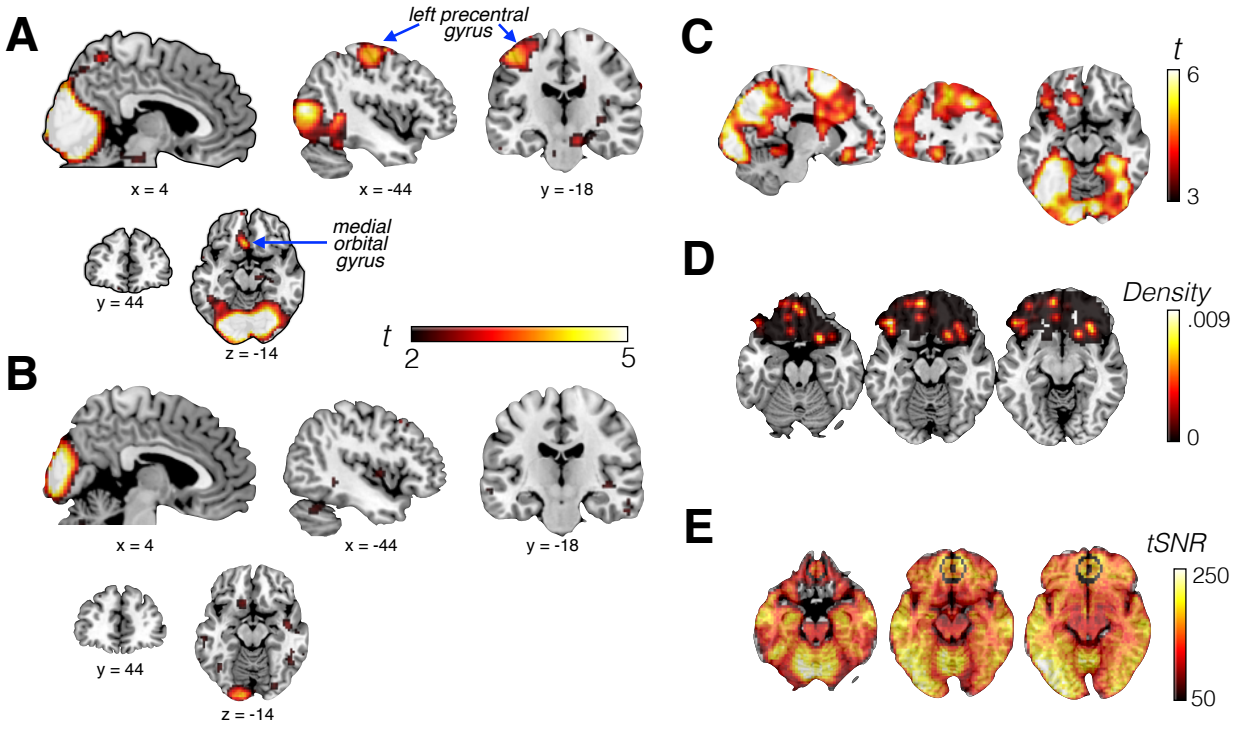

Figure S3. Decoding of motor response and response mapping, distribution of peak effects within OFC and tSNR distribution, related to Figure 3. *(A)*: Whole-brain decoding of motor response (choices were made with the pointing and middle finger of the right hand) showed decoding in visual cortex and left motor cortex. At a lenient threshold, medial orbitofrontal gyrus was also seen. *(B)*: Decoding of response mapping (young=left/old=right or vice versa), in contrast, was only possible in visual cortex, but not motor cortex. Maps in (A) and (B) are thresholded at $t = 2$ for illustration and comparison to Figure 2. *(C)*: Distribution of participant-specific peaks for state-decoding conjunction analysis. Each participant's peak location was convolved with a 3-dimensional gaussian (SD: 3 voxels) and the full distribution normalized to a probability density function, see legend. This analysis was restricted to the anatomical OFC, although the original conjunction analysis was done on the whole brain. *(D)*: Temporal signal to noise ratio (tSNR). Brain maps show color-coded tSNR in different axial slices. The black dot and outline show the result of the conjunction analysis (searchlight center and outline), as reported in Figure 3 of the main manuscript, and do seem to reflect increased SNR in these areas.

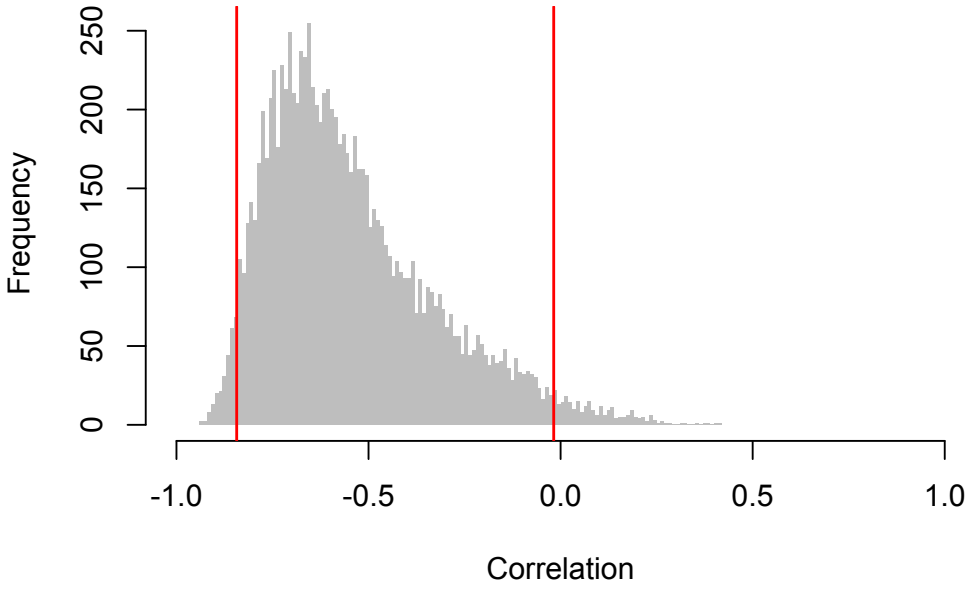

Figure S4. Bootstrapped distribution of correlation coefficients, related to Figure 4 Histogram of correlations as a result of 1000 bootstrapping iterations (sampling with replacement, $n = 27$, i.e. same sample size as original sample). The red lines indicate the 95% confidence interval.

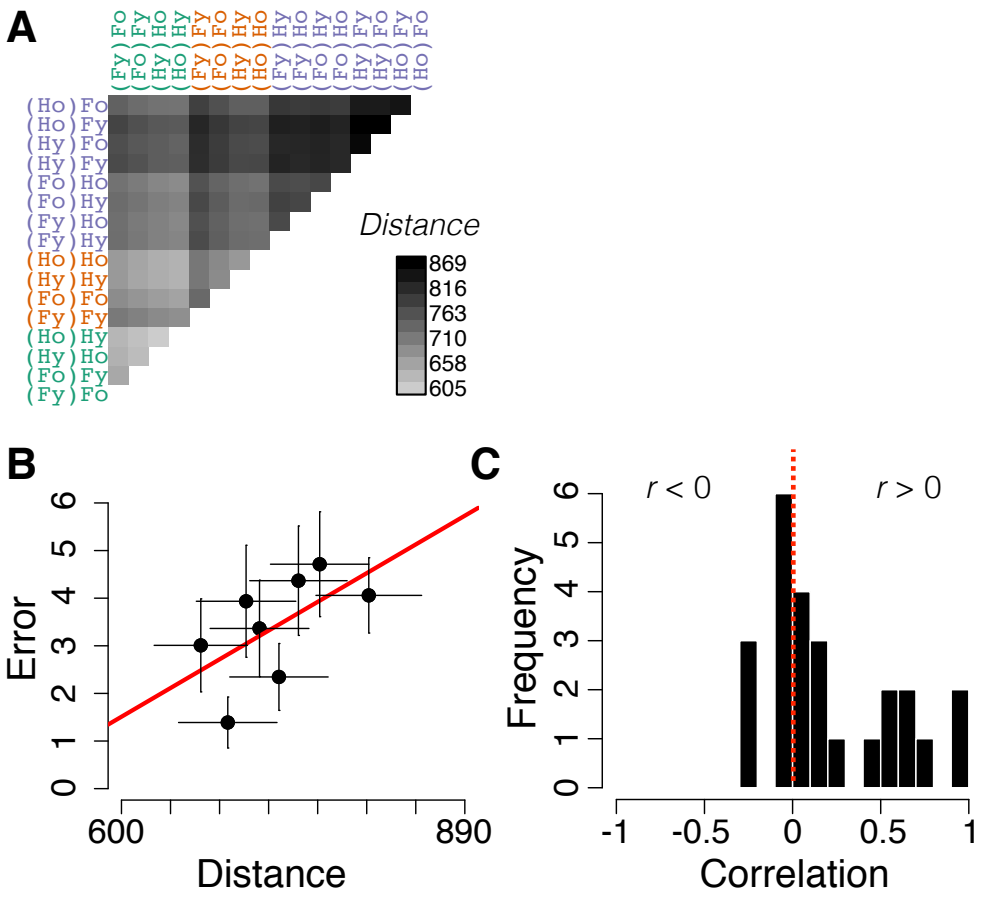

Figure S5. Representational similarity analysis based on between-run Euclidean distances rather than absolute correlations, related to Figure 5. *(A)*: Average euclidean distances between neural state representations within OFC. Darker gray denotes higher distances. *(B)*: Relationship between error rate on the 8 different Exit→Enter transitions and distance between the pairs of states corresponding to these transitions, across participants. Dots denote the average distance between states in that ordinal position across participants (x axis) and average error rate on the corresponding transitions (y axis), with horizontal and vertical error bars denoting S.E.M of each. Lower distance between neural states were associated with fewer behavioral errors, on average ($p = .03$). *(C)*: Histogram of within-subject correlations between error rates and neural state similarity showing that correlations were significantly higher than 0 ($p < .01$).

Table S1. Clusters for wholebrain state component analyses, related to Figure 3. All clusters with $p < .01$ and $k > 15$ are listed. Anatomical names and statistics refer to highest peak of each cluster. Clusters within or extending into OFC highlighted in red.

Anatomical location	Peak (MNI, in mm)			Cluster size	t_{26}	$p_{unc.}$
	x	y	z			
Previous Category						
L Fusiform Gyrus	-42	-34	-20	1046	5.69	$< .0001$
R Precuneus	6	-58	46	986	4.87	$< .0001$
R Middle Occipital Gyrus	33	-73	16	418	4.52	$< .0001$
R Superior Orbital Gyrus	15	47	-14	391	4.50	$< .0001$
R Inferior Occipital Gyrus	42	-64	-14	389	4.09	.0002
L Middle Frontal Gyrus	-33	20	31	223	4.64	$< .0001$
L Superior Frontal Gyrus	-12	38	37	156	3.88	.0003
R SupraMarginal Gyrus	60	-46	43	75	3.28	.0015
L Inferior Frontal Gyrus (*p. orbitalis*)	-30	35	-17	68	4.21	.0001
Previous Age						
R Superior Occipital Gyrus	27	-94	19	91	4.24	.0001
L Superior Orbital Gyrus	-15	56	-2	58	3.38	.0011
L Middle Temporal Gyrus	-45	-55	19	45	3.10	.0022
R Cerebellum	3	-43	-29	36	3.23	.0017
L Superior Parietal Lobule	-24	-70	52	35	3.37	.0012
R Middle Temporal Gyrus	69	-16	-14	26	3.60	.0007
R Inferior Frontal Gyrus (*p. triangularis*)	54	29	22	19	3.00	.0029
R Superior Medial Gyrus	9	44	55	19	3.27	.0015
L Rectal Gyrus	0	44	-14	18	3.56	.0007
Current Category						
L Fusiform Gyrus	-39	-49	-14	14506	9.82	$< .0001$
L Middle Frontal Gyrus	-42	38	22	1733	4.99	$< .0001$
R Inferior Frontal Gyrus (*p. triangularis*)	60	20	19	520	4.60	$< .0001$
R Superior Frontal Gyrus	18	5	61	515	4.45	$< .0001$
L SupraMarginal Gyrus	-63	-40	31	230	5.01	$< .0001$
R Middle Temporal Gyrus	69	-37	7	99	3.52	.0008
L Inferior Frontal Gyrus	-42	32	-14	76	4.11	.0002
R Superior Frontal Gyrus	15	32	46	50	2.86	.0041
L Middle Frontal Gyrus	-42	11	55	23	3.05	.0026
R Precentral Gyrus	48	5	52	20	3.65	.0006
L Middle Temporal Gyrus	-63	-4	-17	16	3.42	.0010
Conjunction						$p_{conjunction}$
R Rectal Gyrus	**3**	**44**	**-14**	**16**	**3.25**	**.0016**

1 Supplemental Results

1.1 Behavioral Results

Behavioral error rates were 2.3% during the main experiment, with no evidence for error rates changing over time (main effect Block: $\chi^2(1) = 0.08, p = .77$). Error rates were not affected by factors Age, Category or their interaction (all p's $> .13$), but were affected by the class of the trial (Enter: 3.4%, Exit: 2.3%, Internal: 1.4%, $\chi^2(2) = 20.8$, $p < .001$). The number of time-outs was negligible (0.3%). The behavioral pretaining and the offer of a performance bonus helped to reduce the number of errors, that is, errors were significantly higher during training than during the main experiment, $t(26) = 5.6, p < .001$. Reaction times (RTs) did not change between training and main experiment (991 vs 985 ms, $t(26) = 0.27$, $p = .79$) nor between blocks within the main experiment ($\chi^2(1) = 1.7$, $p = .18$). RTs were not affected by age ($p = .67$), but were faster for faces than for houses (944 vs 1028 ms, $\chi^2(1) = 78.3$, $p < .001$; no interaction, $p = .98$). As with the error rates, RTs were also affected by trial class with slightly faster trials in Repeat than in the other trial classes ($\chi^2(2) = 153.7$, $p < .001$). Behavioral results are in Figure S1.

The matched error rates for the categories and ages minimized the risk of biases that could confound the results we reported. In addition, to account for the differences between different trial classes, below we present our MVPA analyses separately for Switch (Enter) and Non-Switch (Internal or Exit) trials. Furthermore, we minimized the potential effects of RT differences on decoding results by taking trialwise RTs into account in the first level fMRI analyses (**Todd2013**; **Woolgar2014**). Finally, we investigated potential RT effects on our decoding results by using participants' trialwise RTs in a synthetic fMRI data analysis that simulated the effects of RTs on the BOLD signal in the absence of a true state related neural signal (see Methods; all of these control analyses confirmed our results).

1.2 State Identity Classification

To verify that our ROI-based decoding results within OFC were unbiased, we performed a permutation test by randomly permuting the labels of the training set in each fold, training the classifier in the same manner and assessing its prediction performance in the test set (with unchanged labels). This procedure was repeated 10 times for every participant's data, and the resulting accuracies were averaged within participant. In addition, the contribution of each of the four state components to the 16-way classification based on randomized labels

were assessed in the same manner as in the main analysis. Results showed chance performance for each permutation test (Figure S2A). Specifically, the upper 95th percentiles of the different decoding analyses were all below the classification accuracies obtained with the true data: 7.04% for the 16-way classification, and 51.5%, 51.7%, 51.4% and 51.7% for the four binary comparisons regarding previous category and age, and current category and age, respectively. In addition, we assessed the reliability of the different binary decoding results shown in Figure 2B by calculating bootstrapped distributions (done separately for each of the different state aspects, bootstrapping done over 10000 iterations of sampling participants with replacement).

As an alternative state space for the task, we considered a state definition that included only the three unobservable components previous category, previous age and current category, but not the observable current age, which could be encoded by participants as an action rather than as a state component. This resulted in 8 rather than 16 states. In support of our other analyses, we found that 8-way classification in the OFC was well above chance (16.7%, corresponding to 8.5% above chance baseline, $t_{26} = 6.6$, $p < .001$, see Figure S2C). In line with the results from the 16-state analysis reported in the main manuscript, only OFC allowed the classification of all individual components, but note that past age reached only marginal significance in this analysis (53.5%/$p = .03$, 52.9%/$p = .09$ and 55.5%/$p < .001$, for previous category, previous age and current category, respectively).

1.3 Localization of State Representations

The searchlight-based classification of the four components of the state resulted in four information maps that reflect where in the brain each component could be classified, shown in Figure 3 of the main manuscript (the searchlight analyses followed the same procedure as the main analysis of OFC signals, see Methods). This analysis showed that no part of the OFC encoded the age of the current trial. This could be due to 'current age' being an observable attribute of the state, or it not being an attribute of the state at all. To investigate alternative encoding of action-relevant information, we performed another two whole-brain classification analyses in which we either included the current motor response (Fig S3A) or the current left-right response mapping (Fig S3B) in the state. Specifically, we defined the states according to the current motor response along with the information about the past age and the current and past category (i.e., a state could be defined as '(Fo)Fl', which reflects a trial in which the previous trial was an old Face trial, and the current trial was a Face trial with the correct response being *left*), and similarly for the current mapping.

For the decoding analysis involving current action, the onsets of the trial events in the GLMs were shifted to the onset of the action (in all other analyses, the onsets are at the stimulus onset).

As can be seen in Figure S3A, these analyses showed decoding of the motor action in left motor cortex (responses were made with the index and middle finger of the right hand), as well as visual cortex. At the lenient threshold used for illustration ($T > 2$), a cluster can also be seen in medial orbital gyrus, the same area that showed encoding of previous category, age and current category in the main analysis. However, the effect was detected only at a lenient threshold and was not confirmed in an ROI analysis of the whole OFC ($p = .40$). Moreover, as mentioned above, regressors for the motor action were time-locked to the time of the choice, whereas other state-component regressors were time-locked to stimulus onset. Classification of motor response at the time of the stimulus or of other state components at the time of the response were unsuccessful. Finally, decoding of the current response mapping (whether young was left and old was right, or vice versa) showed mainly decoding in primary visual cortex, but not in left motor cortex (Figure S3B).

We also investigated the spatial specificity of the conjunction effect shown in Figure 3, localized in medial OFC/gyrus rectus. Figure S3C shows the distribution of individual conjunction effect peaks within OFC and indicates rather large across-participants anatomical variability of localization of state representations, which in the case of many subjects involves lateral OFC. We therefore believe that caution is warranted regarding the interpretation of our results pertaining to the precise localization of the state representation within OFC, and do not exclude the possibility that lateral OFC areas are involved in state representations as well. Similarly, the distribution of the temporal signal to noise ratio (tSNR, definition see Methods) indicates a slightly higher SNR in the the medial OFC region which was identified in the group analysis (Figure S3D).

1.4 Correlation between decoding and behavioral errors

To scrutinize the correlation shown in Figure 4, we performed a nonparametric bootstrapping test (1000 iterations, using the R package "boot" with default settings), which confirmed our result (see Figure S4 and main text).

1.5 State Space Similarities

To additionally validate the effect of state representation similarity on error rates reported in Figure 5, we repeated the analyses with (a) Euclidean distances instead of Pearson correlations and (b) simulated data to ensure that the correlations we found were not ascribable to any confounding factors such as temporal proximity or differences in accuracy or reaction times for different trial types (see Methods for procedures used to generate synthetic data). The results confirmed the finding presented in the main text (see Figure S5).