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

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

Princeton Neuroscience Institute and Department of Psychology
Princeton University, Princeton, NJ, 08544, USA

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

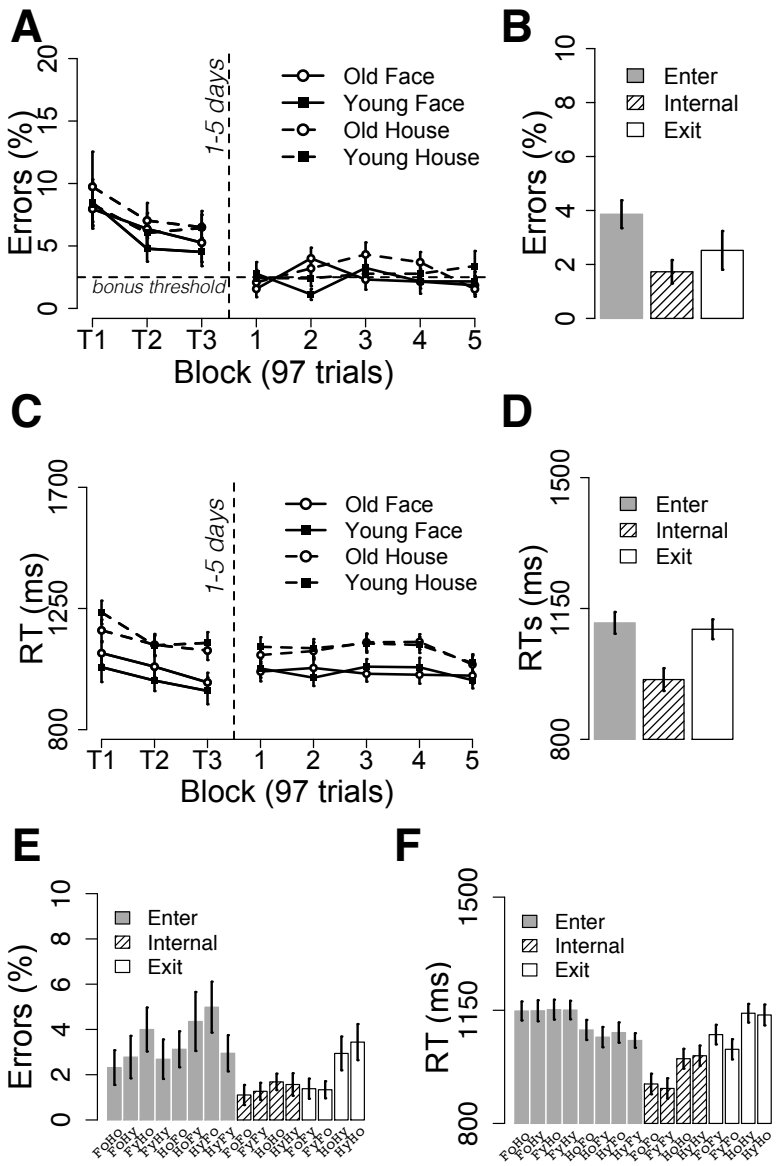

Figure S1. Behavioral results, related to Figure 1. *(A)*: Average behavioral error rate during training and the main experiment. The separate lines distinguish between categories (face = solid lines, house = dashed lines) and the age (filled circles = old, empty circles = young). Dashed horizontal line 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.

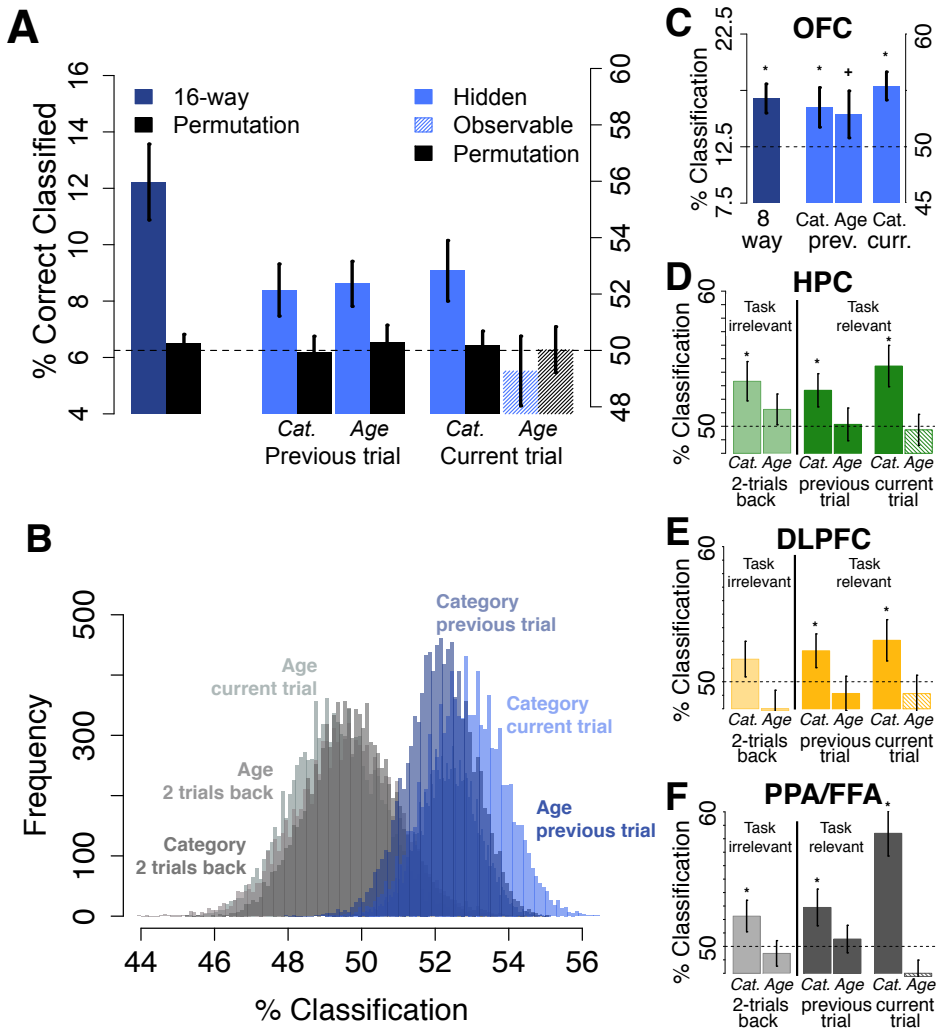

Figure S2. Decoding permutation test and 8-way decoding in different ROIs, related to Figure 2. *(A)*: Results of a permutation test (black bars) show greatly decreased decoding relative to the original decoding shown in Figure 2 (blue bars). *(B)*: Bootstrapped distributions of binary decoding for the 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.

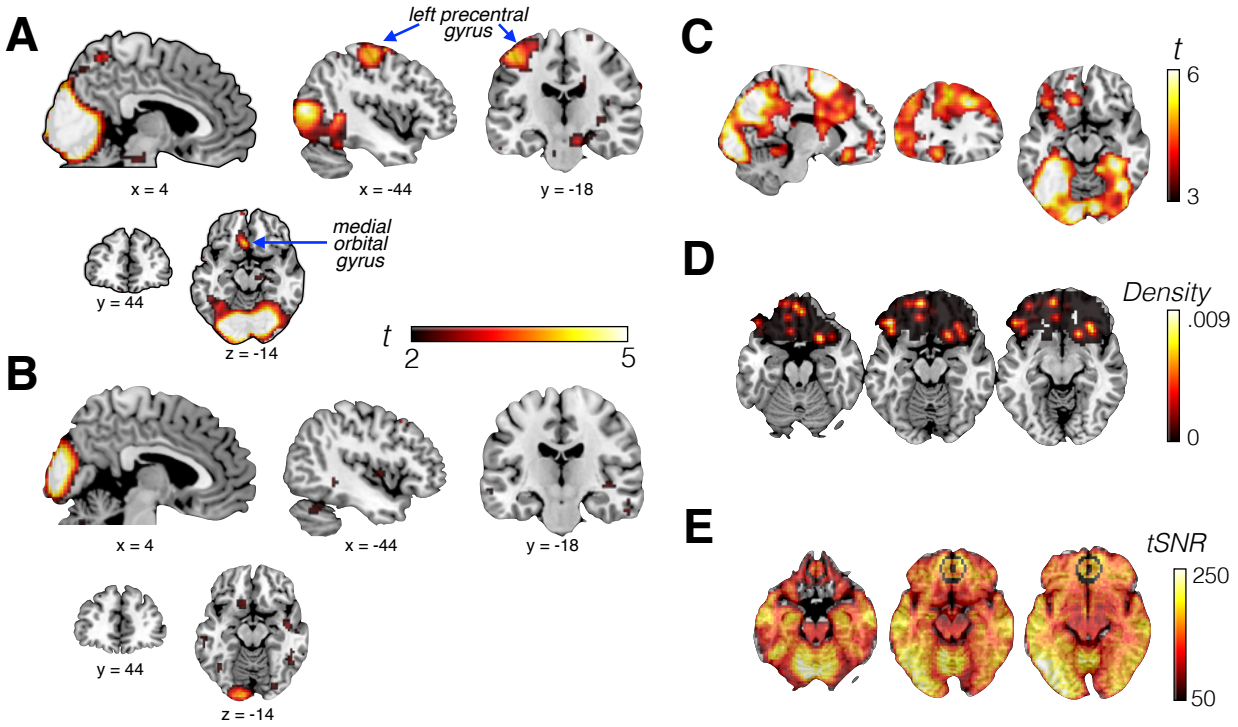

Figure S3. Decoding of motor response and response mapping, distribution of peak effects within OFC and tSNR distribution, related to Figure 3. *(A)*: Whole-brain decoding of motor response (choices were made with the pointing and middle finger of the right hand) showed decoding in visual cortex and 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.

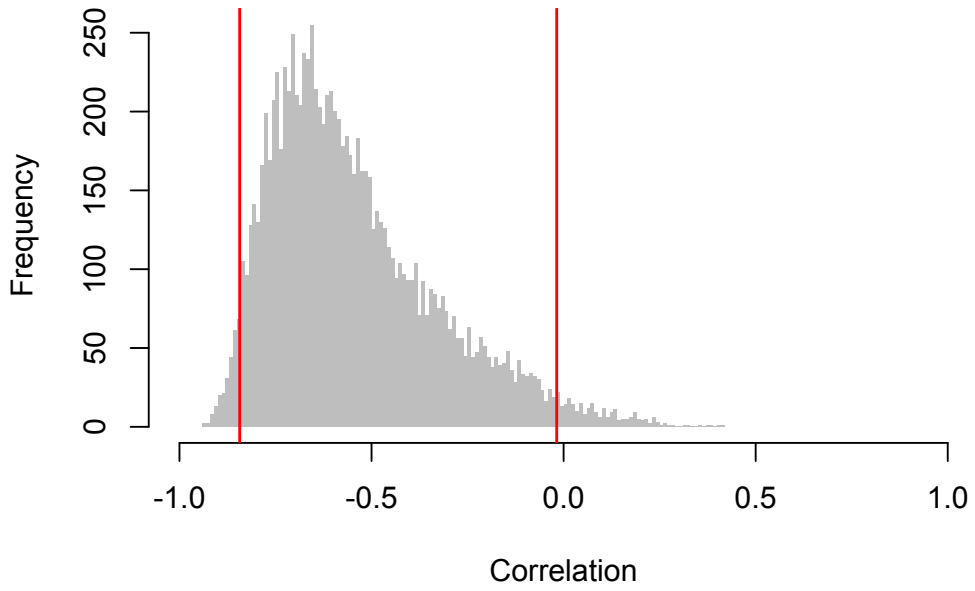

Figure S4. Bootstrapped distribution of correlation coefficients, related to Figure 4 Histogram of correlations as a result of 1000 bootstrapping iterations (sampling with replacement, $n = 27$, i.e. same sample size as original sample). The red lines indicate the 95% confidence interval.

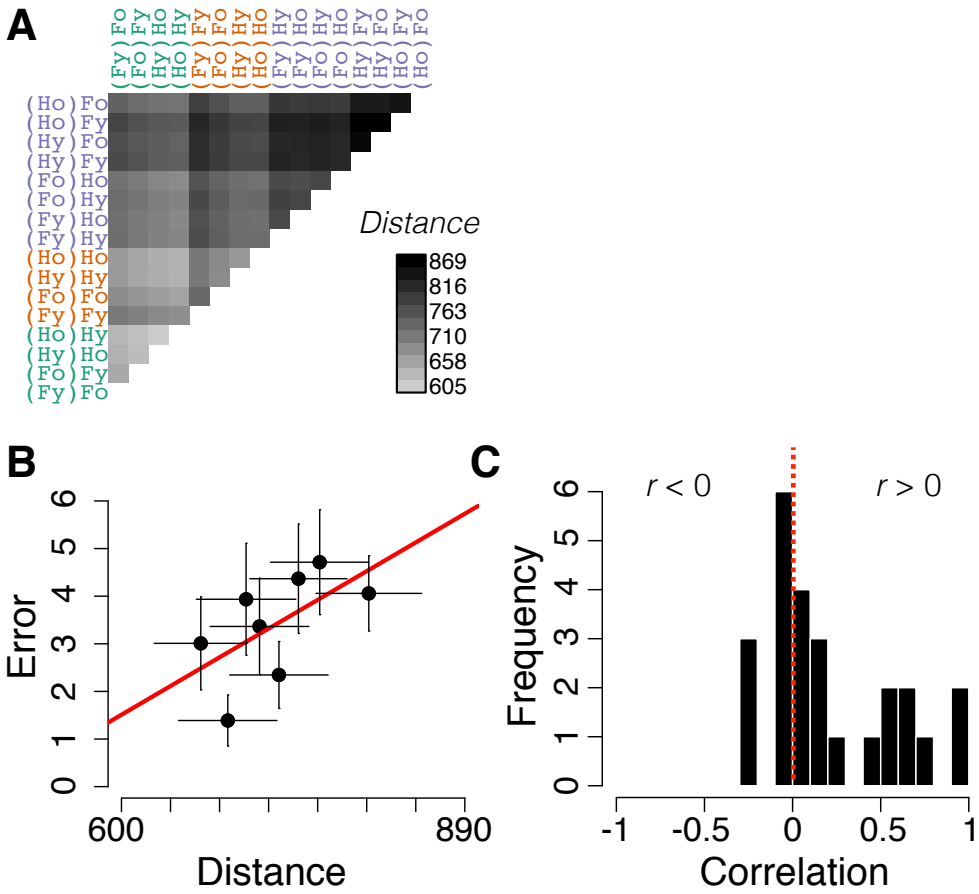

Figure S5. Representational similarity analysis based on between-run Euclidean distances rather than absolute correlations, related to Figure 5. *(A)*: Average euclidean distances between neural state representations within OFC. Darker gray denotes higher distances. *(B)*: Relationship between error 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) x	y	z	Cluster size	t_{26}	$p_{unc.}$
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).