Supplementary Materials for

Sequential Replay of Non-spatial Task States in the Human Hippocampus

Nicolas W. Schuck[1,2,3,*] & Yael Niv[3,*]

[1]Max Planck Research Group NeuroCode, Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany
[2]Max Planck UCL Centre for Computational Psychiatry and Ageing Research
Berlin, Germany & London, UK
[3]Princeton Neuroscience Institute and Department of Psychology, Princeton University
Washington Road, Princeton, NJ, 08544, USA

Correspondence to: schuck@mpib-berlin.mpg.de, yael@princeton.edu

This PDF file includes:

Materials and Methods

Supplementary Text

Figs. S1 to S7

Tables S1 and S2

Materials and Methods

Participants

Thirty nine participants were selected according to standard fMRI screening criteria (right handedness, 18–35 years of age, normal or corrected-to-normal vision and no contraindication for fMRI) from the Princeton University community, and were compensated with $20 per hour plus up to $5 performance-related bonus. Six participants were excluded from analysis due to either technical errors (3 participants), violation of performance criteria standards (2 participants with over 4 times the average error rate in the last two blocks of the experiment) or incomplete data (1 participant). The final sample consisted of 33 participants (22 female, mean age 23.4 years). All participants provided informed consent and the study was approved by Princeton University's Institutional Review Board.

Stimuli

Stimuli consisted of spatially superimposed images of a face and a house (see *(7)*; face images from `http://faces.mpdl.mpg.de/faces` described in *(53)*, see Fig. 1 Main Text). Faces and houses could be classified as either young or old, e.g. a stimulus could show an old face image blended with a contemporary (i.e., young) house image. Four classes of stimuli were possible: (1) two old or (2) two young face and house pictures, (3) a young face with an old house or (4) vice versa.

Task

The task was identical to Schuck et al. (2016) and will be described only briefly. Each trial began with the display of the mapping of a left and right key to a young and old response (changing randomly trialwise) below a fixation cross for 1.2s (range: 0.5–3.5s). Then, a compound face-house stimulus was shown for 3.3s (range: 2.75-5s; Main text Fig. 1) and participants had to

make an age judgment about one of the two image categories. Participants knew which category of the stimulus they had to judge by applying the following rules: 1. Before the first trial of each run, the category to judge was displayed on the screen; 2. Once the age of the relevant category changed (e.g., from young to old), the judged category changed on the next trial. 3. No age comparison was necessary on the first trial after a category change, i.e. each category was judged for at least two trials in a row before a switch. The average trial duration was 4.5s (range: 3.25-8.5s), all timings were randomly drawn from a truncated exponential distribution and the response deadline was 2.75s. The category instruction cue at the beginning of a run was displayed for 4s. Erroneous or time-out responses led to feedback (written above stimulus for 0.7s) and trial repetition. If an error trial involved an age change (and thus would require a category switch on the next trial), participants had to repeat the trial before the error as well as the error trial, giving them the chance to observe the age change. Otherwise, they had to repeat the trial on which they made the error.

Design

Participants underwent two fMRI sessions. The first session began with the display of written instructions while participants underwent a functional scan (group 1), or a 5 minute resting-state scan followed by instructions (group 2). The instructions explained the rules of the task and contained a training phase in which simple age judgments had to be made on (non-overlapping) face and house images. The images shown in this period were later used in the task, thus familiarizing participants with the age judgment aspect of the task as well as the stimuli. The instructions furthermore involved an annotated walk-through of four trials of the real task (i.e., with overlapping images and the requirement to switch attention after an age change). Following the instructions, participants performed 4 runs of the task (97 trials per run, 388 total). Each run lasted about 7-10 minutes and participants were given the chance to rest briefly between

runs. A 5 minute fieldmap scan was done between runs 2 and 3, resulting in a longer break for participants. After run 4, participants underwent a resting state scan as well as a structural scan. Lights were turned off during resting-state scans and participants were instructed to stay awake for the entire duration of the scan (5 minutes, 100 TRs). The second session was identical for all participants and involved the following scans: resting state, 2 task runs, fieldmap, 2 task runs, resting state and structural scan. Thus, overall, participants performed 8 task runs and 3 (group 1) or 4 (group 2) resting-state scans. In all other regards, the task design involved the same characteristics as detailed in Schuck et al. (2016).

Behavioral Analyses

Behavioral analyses were done using mixed effects models implemented in the package lme4 *(48)* in R *(49)*. The model included fixed effects for Block, Condition, Category and intercept. Participants were considered a random effect on the intercept and the slopes of all fixed effects. The reported p-values correspond to Wald chi-square (χ^2) tests as implemented in R. Reaction time (RT) analyses were done on error-free trials only and reflect the median RT within each factor cell.

fMRI Scanning Protocol

Magnetic-resonance images were acquired using a 3-Tesla Siemens Prisma MRI scanner (Siemens, Erlangen, Germany) located at the Princeton Neuroscience Institute. A T2*- weighted echo-planar imaging (EPI) pulse sequence was used for functional imaging (2×2 mm in plane resolution, TR = 3000 ms, TE = 27 ms, slice thickness = 2 mm, 53 slices, 96×96 matrix (FOV = 192 mm), iPAT factor: 3, flip angle = 80°, A→P phase encoding direction). Slice orientation was tilted 30° backwards relative to the anterior – posterior commissure axis to improve acquisition of data from the orbitofrontal cortex *(46)*. Field maps for distortion correction were acquired using the same resolution (TE1 = 3.99ms), and a MPRAGE pulse sequence was used

to acquire T1-weighted images (voxel size = 0.9^3 mm). The experiment began 20 seconds after acquisition of the first volume of each run to avoid partial saturation effects.

fMRI Data Preprocessing

FMRI data preprocessing was done using SPM8 (`http://www.fil.ion.ucl.ac.uk/spm`) and involved fieldmap correction, realignment, and co-registration to the segmented structural images. The task data used to train the classifier were further submitted to a mass-univariate general linear model that involved run-wise regressors for each state (see below), nuisance regressors that reflected participant movement (6 regressors) and run-wise intercepts. The resulting voxelwise parameter estimates were z-scored and spatially smoothed (4 mm FWHM). The resulting activation maps were used as the training set for a support-vector machine (SVM) with a radial basis function (RBF) kernel using LIBSVM *(51)*. The SVM was trained to predict the task state from which a particular activation pattern came from. Like the activation maps used for classifier training, the resting-state data were z-scored and smoothed (4mm FWHM). Anatomical ROIs were created using SPM's `wfupick` toolbox. The hippocampus (HC) was defined as the left and right hippocampus AAL labels. The orbitofrontal cortex was defined as in *(14)*. Behavioral analyses and computations within the assumed graphical model of state space (see below) were done using R *(49, 52)*.

Autonomic nervous system activity during task and rest

In order to assess whether there were any differences in vigilance or arousal between the different resting conditions (and the task), we analyzed the data recorded during the experiment from the Siemens MRI optical pulse sensor and pneumatic respiratory belt. We then used the TAPAS PhysioIO toolbox *(50)* to determine participants' heart rate.

Data was successfully recorded in at least one session in 30 out of 33 participants. The average heart rate across the entire experiment was 69.7 beats per minute (SD: 10.4). As expected,

this data showed that participants had generally higher heart rates during the task compared to the average rest conditions, $t_{29} = 6.2$, $p < .01$, presumably due to heightened vigilance and arousal (mean difference: 3.8 beats per minute more during the task). Importantly, however, there was no difference between the pre (PRE + INSTR) and the post rest conditions (all TASK rest conditions), $t_{22} = 0.7$, $p = .43$ (differences in df's arise from physiological recording failures in some participants/sessions). In addition, there was no difference between the PRE condition and any of the TASK rest conditions (ps = .23, .96 and .75 for the 1st, 2nd and 3rd resting state, respectively), between the INSTR and any of the TASK rest conditions (ps = .11, .65 and .20, respectively), or between the PRE and the INSTR conditions ($p = .15$). In addition, there was no relation between heart rate during the TASK rest condition and sequenceness effects during the TASK rest condition, $r = -0.05$, $p = .79$.

fMRI Classification Analysis

The support vector machines were trained on 8 maps of parameter estimates ("betas") for each of the 16 states (one map for each state and run) restricted to the anatomical mask of the hippocampus (back-transformed into each participant's individual brain space) or the orbitofrontal cortex. Classification accuracy was assessed with a leave-one-run-out cross-validation scheme in which data from 7 runs were used for training and the held-out run was used for testing (Main Text; Fig. 2). The resting-state analysis used a classifier trained on all available task data (8 runs). This classifier was applied to each volume of the resting-state data and the most highly classified state was considered as the output of the classifier for that volume. The resulting sequence of predictions was the main focus of our analyses (see below). We obtained the distance to the hyperplane by dividing the decision value by the norm of the weight vector w, as specified in the libSVM webpage (http://www.csie.ntu.edu.tw/ cjlin/libsvm/faq.htmlf4151). For each volume, we then calculated the average of the distances of all pairwise comparisons of

the predicted class against all other classes, to obtain a proxy of how certain the classifier is in its prediction. Student t tests pertaining to decoding results were one-tailed, given the *a priori* expectation of larger-than-chance decoding in the hippocampus.

Sequenceness Analysis

The main question of the sequenceness analysis was whether the state transitions decoded from resting state scans, T, were related to the distance between states experienced during the task, D. To this end, we analyzed the neural state transitions T with logistic mixed-effects models, using the lme4 *(48)* package in R *(49)*. Because the slow hemodynamic response function leads to encoding of sequential structure in activity patterns (i.e., there is high similarity between temporally adjacent patterns), a classifier trained on sequential task data can be biased to decode states in a similar sequence to the training data, even if the test data are random (i.e., the 'sequenceness' identified in the test data comes from the training data, not the test data). We therefore applied the trained classifier to matched fMRI noise (see below) and used the resulting spurious 'state transitions,' $T[\epsilon]$, as a covariate that would account of the spurious base rate of transitions that is due to the classifier rather than the data. Applying these models to control conditions consistently yielded non-significant effects of sequenceness, showing that this analysis appropriately controls for the above mentioned spurious structure that is observable for instance in the significant correlations between D and T in the noise data (Main Text; Fig. 3F). Specifically, our model included the following fixed effects: (1) the distance between states, D, which was the regressor of interest, and as regressors of no interest (2) the transition probabilities obtained in the above mentioned noise simulations, $T[\epsilon]$, (3) an orthogonal quadratic polynomial of $T[\epsilon]$ that was included in order to account for as much noise-related variance as possible, and (4) an intercept. Models of change in sequenceness across PRE, INSTR and TASK conditions (Main Text; Fig. 4) additionally involved interaction terms of condition with the distance D and

condition with the noise transitions $T[\epsilon]$. Participant identity was included as a random factor to account for between-subject variability. To capture state-related variability (state frequency effects affect the distribution of state transitions), state identity s_j of a transition from state i to state j was used as an additional random effect nested within subject. Participant and state were random grouping factors for all fixed effects with exception of the quadratic expansion of $T[\epsilon]$, where including these random factors caused problems in fitting the logistic regression models.

Formally, the model followed the general assumption that the number of transitions Y is drawn from a binomial distribution of n draws and probability T:

$$Y_{ijk} \sim B(n_k, T_{ijk})$$

where n_k corresponds to the number of measurements for subject k, and i and j index the outgoing and incoming states of a given transition. The logit transformed probabilities T (shown in Main Text Fig. 2D; logit is the canonical link function for binomial models) were then modeled in a mixed effects regression model with the above mentioned fixed and random effects structure:

$$\begin{aligned} logit(T_{ijk}) &= \beta_0 + D_{ij}\beta_1 + T[\epsilon]_{ij}\beta_2 + T[\epsilon]^2_{ij}\beta_3 + \\ &\quad \gamma_{0k} + D_{ij}\gamma_{1k} + T[\epsilon]_{ij}\gamma_{2k} + \\ &\quad \zeta_{0kj} + D_{ij}\zeta_{1kj} + T[\epsilon]_{ij}\zeta_{2kj} + \epsilon_{ijk} \end{aligned}$$

In the text, we describe the fixed effect of D, β_1 in the models, as 'sequenceness,' and the fixed effect of $T[\epsilon]$, β_2, as 'randomness' (Main Text; Fig.s 4B,C). The subject-specific random effects of D, γ_{1k}, were used as individual sequenceness indicators in the correlations (Main Text; Fig.s 4F,G). The state- and subject-specific random effects are indicated by ζ. Correlations between random effects were estimated. Model comparisons were conducted using likelihood-ratio tests by comparing models that included the noise transitions $T[\epsilon]$ with versus without the

fixed effect regressor of distance (sequenceness), or without the condition interaction terms to the full models that included these terms. The random effects structure was kept constant across these comparisons.

T-tests pertaining to sequenceness results (number of steps, etc.) are one-tailed, given our *a priori* expectation of larger sequenceness in the hippocampus compared to the various controls.

Alternative Task Transition Matrices

As mentioned in the main text, for reasons of model comparability we used the 1-step transitions of the task as a basis to test alternative replay models. The 1-step transition matrix simply reflects from which state one could proceed to which other states in one trial. The alternative task transition matrices were based on the assumption that the hippocampus has access to only partial state information, and hence correspond to transition matrices defined over subsets of states.

For instance, to compute the transition matrix of the "stimulus model" we defined $\mathcal{S}_{Fy}^{stim}$ as the subset of states in which Fy was the stimulus:

$$\mathcal{S}_{Fo}^{stim} = \{(Fy)Fy, (Fo)Fy, (Hy)Fy, (Ho)Fy\}.$$

$\mathcal{S}_{Fo}^{stim}$, $\mathcal{S}_{Hy}^{stim}$, $\mathcal{S}_{Ho}^{stim}$ were the corresponding subsets of states in which Fo, Hy and Ho were the stimuli, respectively. The 1-step distance matrix was then computed such that every transition between two states s_i and s_j in the complete task state diagram was converted into four transitions from s_i to all four states that are part of the same subset as s_j, that is $\mathcal{S}_j^{stim}$. All resulting transitions are summed, and normalized so that all exiting transitions from a state would sum to 1. The new transition between $\mathcal{S}_{Fy}^{stim}$ and $\mathcal{S}_{Fo}^{stim}$ would therefore count within it $(Fy)Fy \rightarrow (Fy)Fo$, $(Ho)Fy \rightarrow (Fy)Fo$ and $(Hy)Fy \rightarrow (Fy)Fo$, whereas the new transition between $\mathcal{S}_{Fy}^{stim}$ and $\mathcal{S}_{Ho}^{stim}$ would only count within it $(Fo)Fy \rightarrow (Fy)Ho$. After normalization, these would be $3/8$ and $1/8$, respectively.

For the other alternative models, we defined subsets of states that have the same current attended category, and subsets of states that have the same current and previous attended categories, and then computed the transition matrices as described above. The 1-step transition matrices of these alternative models are shown in Figures 5A-C in the Main Text. The reverse replay transition matrix was simply the transpose of the full task 1-step transition matrix.

Synthetic fMRI Data and Noise Simulations

In order to estimate to what extend training the classifiers on sequential data influenced the sequenceness of their predictions, we simulated, for each participant, individually spatio-temporally matched fMRI noise, and applied the classifiers to these data. For each participant and resting state session, we first extracted fMRI data from the hippocampus and the orbitofrontal cortex and calculated the voxel wise mean of that participants' real resting state data as a baseline for each TR. We then added temporal noise to this baseline as described below, ensuring that spatial properties driven by the anatomy and tissue partial volume are reflected in the noise data used for the simulations (i.e., the average noise data looks like each participants' real data reflecting grey/white matter differences etc.).

As in the classification analyses, we applied linear detrending to each voxel. We then estimated the average standard deviation of the voxels within these regions, as well as the average autocorrelation using an AR(1) model in R. Next, we used the neuRosim toolbox in R *(52)* to simulate fMRI noise with the same standard deviation and temporal autocorrelation as the real data. Finally, we used AFNI's 3dFWHMx and 3dBlurToFWHM functions to first estimate the spatial smoothness of the real data, and then smooth the simulated noise until it had the same effective smoothness.

For each existing resting-state run, matched noise data with the same number of TRs and voxels were generated. Figure S1 shows the temporal and spatial smoothness of the real and

simulated data separately for each run. In all cases, the properties of the simulated data did not differ from the real data, paired t-tests, all ps $> .05$.

Finally, we applied each participant's classifier to the matched noise data. The classifier was identical to the classifier that was used in estimating the sequences of states from the real data. The resulting sequence of predicted states reflects the bias of the classifier to make sequential predictions because of pattern overlap in the training set, even when applied to noise, as well as any tendency of the classifier to decode certain states more often than others. We therefore used the sequence of states from this analysis to construct the nuisance covariate for the mixed effects models, i.e. the noise 'transition matrix', $T[\epsilon]$, and to perform the appropriate comparisons in the correlation analysis. The average noise transition matrix is shown in Fig. S2. These comparisons between sequenceness in real data versus simulated noise in the correlation and mixed effect analyses indicated that the noise sequenceness $T[\epsilon]$ indeed explained a significant amount of sequential variability of the decoded states (see Main Text Figs. 3F,G, 4B, D), and thus served as a powerful control. Together with the tests done on permuted data (Main Text; Fig. 3B-D, 3F,G), the comparisons across brain regions (Main Text; Fig. 4E) and the within-participant comparisons between the PRE, INSTR and TASK conditions (3B-D, H and 4A-D), our approach represents a stringent control of potential biases.

Transition exclusion analysis

We assessed the effect of temporal contingencies in the data used to train the classifier on the sequenceness measure in the resting state scans by excluding certain transitions from the data used for classification training. Specifically, for all states that had more than one incoming transition (blue states in Figure S3), we excluded from the classifier training those trials in which these states were preceded by one of the preceding states (see Fig. S3: trials corresponding to black arrows were included, trials corresponding to red arrows were not). In two separate

sets of classifiers, we swapped which of the transitions were excluded as shown in Fig S3. For instance, the state (Hy)Ho was preceded in 25% of the trials by the state (Fo)Hy, and in 25% by (Fy)Hy (and in 50% by (Hy)Hy). We therefore trained two separate classifiers, one that excluded the (Fo)Hy $\rightarrow$ (Hy)Ho transitions, and one that excluded the (Fy)Hy $\rightarrow$ (Hy)Ho transitions. The sequential decoding analysis was then performed on the resting scans with both classifiers, and we compared how often the excluded versus the matched included transitions were observed during rest. We took the non-left out transitions ending in the same state as the matched comparison, such that the number of states, the possible more-than-one step transitions between them, and the involved terminal state were equated. If pattern similarity in the classifier training set determines the frequency of decoded transitions, then excluded transitions should be decoded less often in the resting scans as compared to included transitions. If, in contrast, sequential replay has a significant contribution to the observed transitions during rest, the excluded transitions should be observed equally often, or at least more often than transitions that correspond to a larger distance in the the state space.

Replay simulations

In order to assess the sensitivity of our analysis to effects of fast sequences of neural events on fMRI data, we simulated 'neural' replay events at different speeds and applied our analysis to the simulated data. Each simulation involved 240 'units', which generated neural activity on the order of milliseconds. Each unit drove the activity of one voxel, the observational unit in fMRI (Fig. S4A). Activity in the units was a mixture of signal and noise. The noise component was drawn from a multivariate Gaussian distribution with a diagonal covariance band that reflected spatial autocorrelations between neighboring units. The signal reflected sequential activation of patterns which represented one of 16 possible states. Specifically, the replay was simulated as a fast sequential activation of the involved states. Each state was activated for 2 ms, and involved

14 units. Ten of these units were unique to this state, and 4 units were non-unique, that is they were also activated in response to at least one other state. Neural activity corresponding to the activation of a sequence of states could then be simulated by activating the state patterns in the given order (Fig. S4B). The strength of activation caused by the patterns relative to the background noise was treated as a parameter that was systematically varied across simulations, and unit-wise deviations of this average activation strength were drawn from a Gaussian distribution with a standard deviation of 0.5. The strength of the average activation of units during the representation of a state reflects the signal-to-noise-ratio (SNR) in the simulated data and affects decoding accuracy (see below). Activity in units was converted to activity in voxels through convolution with the canonical double-gamma hemodynamic response function:

$$h(t) = \frac{t^5 e^{-t}}{\Gamma(6)} - \frac{t^{15} e^{-t}}{6\Gamma(16)}$$

where t references time and Γ represents the gamma function. The resulting temporo-spatial activation pattern (Fig. S4C) was then used to train and test the classifier as described below.

With this general setup, we generated a 'task' data set and a 'rest' data set. In the task data, the sequences of states were activated in sequential order (all combinations, where face always comes before house and young before old, i.e. FyFy, FyFo, FyHy, FyHo, FoFy, ...) with the same temporal spacing as in the task. The task data set contained 8 runs in each of which every state was activated once. Note that, unlike in the real data, the order of state activation in the simulated task data was unrelated to the order of replay during rest. This ensures that the effects found cannot be explained by a bias for sequentiality in the classifier due to the training data set.

In addition to the task data set, we generated 300 TRs of rest data. During rest, noise was drawn from the same multivariate Gaussian used for the task data simulation (units exhibited the same covariance structure). Replay events were assumed to occur spontaneously every 30

seconds during rest *(12)*, with inter-replay-intervals jittered according to an exponential distribution (intervals were defined as the period from end of the previous sequence to the beginning of the next sequence). Each replay sequence was between 2 and 10 states long, following an exponential distribution *(3)*. Replayed states followed the same sequence as the one used in our main task (but different from the one used to activate states in the simulated task data). The temporal delay between state activations (i.e., the replay speed) was treated as a second parameter, whose effect on the results will be investigated along with the effect of the SNR.

The resulting unit time-courses were convolved with the HRF as described above, and subsampled to provide one 'volume' of brain activity consisting of voxels sampled every 2.9 seconds as in our real data. The subsampled voxel time-courses were then z scored. As in the analysis presented in the main manuscript, we then trained a classifier on all task data and applied it to every volume of the resting state data. Finally, we analyzed the resulting sequence of classifier predictions as in the real data via correlations between pattern transition frequency and state distance of the underlying replayed sequences.

Supplemental Text

Supplemental Results

Behavior

Figure S5 shows reaction times (RTs) and error rates separately for each state and session. As reported in the text, error rates and RTs decreased over time, and were also affected by the state type.

Mixed effect models: Alternative distance metric

To test for the generality of our results, we repeated the model comparisons reported in the main text using a different distance metric for task states. Recall that in the main analysis

we computed distance between each pair of states as the minimum number of steps needed to traverse from one state to the other, in the state space. Here, instead, we simulated a random walk within the task's state space, and obtained the *average* number of steps it took to tranverse between two states. This metric was taken as the distance metric D instead of the minimum number of steps. With this new D we were able to replicate all the results from the mixed effects analysis reported in the main text. In particular, we again found a significant interaction between distance and resting state scan timepoint (TASK vs PRE) in modeling the empirical transition matrix T (significantly better fit when the distance by timepoint interaction term was included, $\chi^2_3 = 11.19$, $p = 0.01$), suggesting that there was a significant difference in the relationship between T and state distance when comparing the resting scan before session 1 with the scan after session 1. Below is a table listing the fixed effects, AICs and whether a likelihood ratio model comparision test was significant for all models reported in the main text.

Mixed effect models with OFC data

We also repeated the mixed effects models reported in the main text using data from the orbitofrontal cortex instead of the hippocampus. The table below lists all results. Importantly, we did not find any evidence in favor of sequenceness increases from PRE to TASK, or from INSTR to TASK.

Replay simulations

Using simulations we tested whether, in principle, rapid replay events can be reliably decoded from BOLD data, given their low temporal resolution and correlated noise. To this end, we simulated replay events in resting data (see Materials and Methods) and attempted to decode them with classifiers trained on simulated task data. Our simulations showed that indeed classifier predictions during rest could be related to replay events. In a simulation in which decodability of states was high (SNR: 15) and replay speed was slow (temporal delays between replayed

states: 2560 ms), a simulated replay event could be decoded quite reliably in the volumes following the event (Fig. S6A). More interestingly, even when decoding success was realistically low (only $\sim 12\%$, as in our experiment) and replay events were very fast (40 ms between simulated pattern activations), the classifier was able to decode the states belonging to the replayed sequences in the TRs following the replay event (Fig. S6B, S6C). This indicates that the superposition of state-related patterns caused by rapid replay events does not hamper the classifiers' ability to detect the presence of states.

Notably, the decoding accuracy for replay events was similar to the decoding of individual, well-separated states. This effect has important implications for the feasibility of measuring replay events in fMRI, where classification accuracy for our 16-state task is low. Let's call the decoding accuracy of 11.6% in the hippocampus, which reflects task conditions in which a volume is characterized by a single true state, a. In the volumes following a replay event, several states impact the BOLD signal. That each of these states is decodable with a similar accuracy a (see Fig. S6B) increases our ability to observe sequential events in two consecutive volumes. In particular, the chance that the states decoded in two consecutive TRs following a replay of, say, 5 steps, are in the correct order (and therefore result in a low distance measure that we use in several of our analyses in the main text) is the sum of the probability that the first TR is the first state of the replayed sequence, times the probability that the prediction on the next TR is any of the states 2-4, plus the remaining possibilities if the first decoded states are later in the sequence:

$$\begin{aligned} P(ordered\,pair) \sim\; & P(C_1 = s_1)P(C_2 = \{s_2, s_3, s_4, s_5\}) \\ & + P(C_1 = s_2)P(C_2 = \{s_3, s_4, s_5\}) \\ & + P(C_1 = s_3)P(C_2 = \{s_4, s_5\}) \\ & + P(C_1 = s_4)P(C_2 = s_5) \end{aligned} \tag{1}$$

where $P(C_1 = s_1)$ reflects the probability that the decoded class C in the first volume is the first state of the sequence s_1, and $P(C_2 = \{s_2, s_3, s_4, s_5\})$ reflects the probability that any of the states 2-5 is decoded on the second volume. Because in the volumes following a replay event all involved states independently have roughly the same chance to be classified as in the regular classification analysis, we can substitute a for each probability of a single state, and $n \cdot a$ for the probability to classify *any* of n states, giving

$$P(ordered\,pair) \sim 4a^2 + 3a^2 + 2a^2 + a^2. \tag{2}$$

For a given replayed sequence of length l, the probability of an ordered pair can therefore be written generally as

$$P(ordered\,pair) \sim \sum_{i}^{l-1}(l-i)a^2 = a^2\frac{l^2-l}{2} \tag{3}$$

Although this relationship only represents a back-of-the-envelope calculation, it illustrates that the probability to observe an ordered pair is significantly higher than a^2, the intuitive estimate of observing two correctly classified states in a row, assuming the decoding accuracy on task data. For the example of a replayed sequence of $l = 5$ steps, and a decoding accuracy of $a = 0.116$, as in our case, the probability of observing an ordered pair is approximately $0.116^2(\frac{25-5}{2}) = 0.13456$, i.e. about 13.5%.

Indeed, the superposition of multiple patterns triggered by the replay lasts for several volumes. Even very short replay events therefore allow the recovery of a substantial proportion of states involved in the replay event. Interestingly, our simulations show that overall decoding accuracy in each volume is equally high under fast compared to slower replay events (Fig. S7A). Even though under low SNR conditions the order of predicted states following a fast replay event does not necessarily reflect the order of replayed states, the classifier will decode states which come from the same *sequence*. This leads to a lower average number of steps between

decoded states following a replay event than expected by chance (Fig. S7B), offering a plausible explanation for the findings reported in the main text (Fig. 3C,D). In other words: because the replayed states are close by in state space, even a random ordering of these same states will be more sequential than expected when states are drawn at random from the entire state space. This effect also leads to the correlation between D and T that we report in the main paper (Fig. 3F-H), which we find to be detectable at 12% decoding success already for replay speeds of about 40-60ms (Fig. S7C,D).

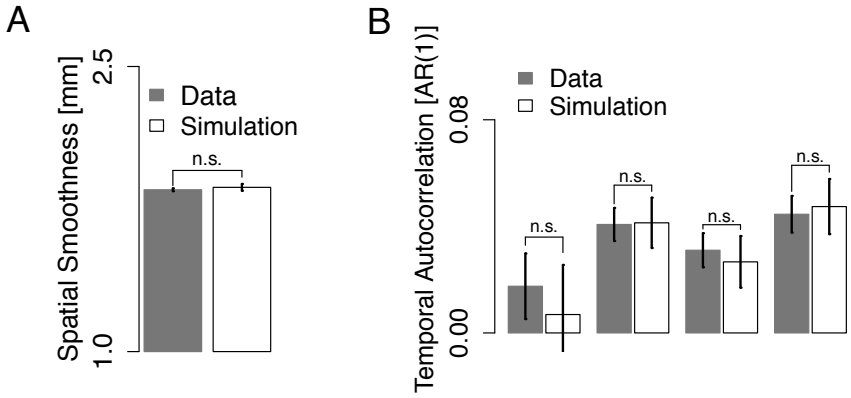

Figure S1. Spatial and temporal properties of real data and noise. *(A)*: Estimated spatial smoothness in simulated fMRI noise (white bar) and real data (grey bar). Smoothness reflects the average full width at half maximum of a gaussian function across x, y and z directions. *(B)*: Estimated autocorrelation coefficients in simulated fMRI noise and real data, separately for each resting scan. Coefficients were estimated using a standard AR(1) model. Error bars represent $\pm$ SEM, n.s: $p > .05$.

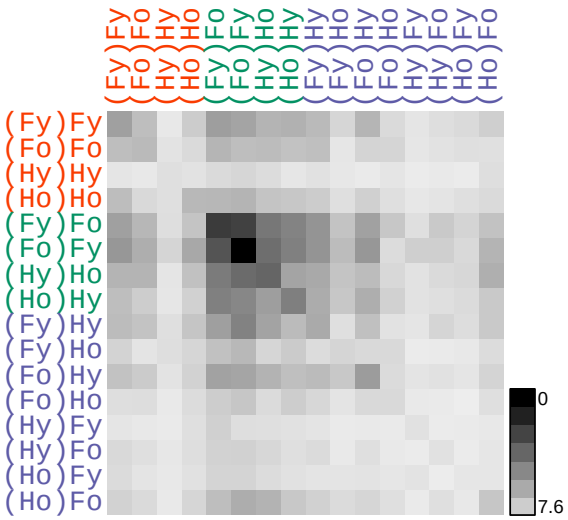

Figure S2. Spurious state transitions estimated in synthetic data. The matrix shows the (non-normalized) number of transitions between states found when the classifier was applied to matched noise data (see Main Text and SI). These results reflect the bias of the used classifiers to report certain transitions more often than others. Individual estimates per participant and conditions were used in the mixed effects analysis as the nuisance regressor T[ϵ], see Main Text. Values shown in Figure reflect the average across all participants and all POST conditions (see color scale, lighter = more frequent).

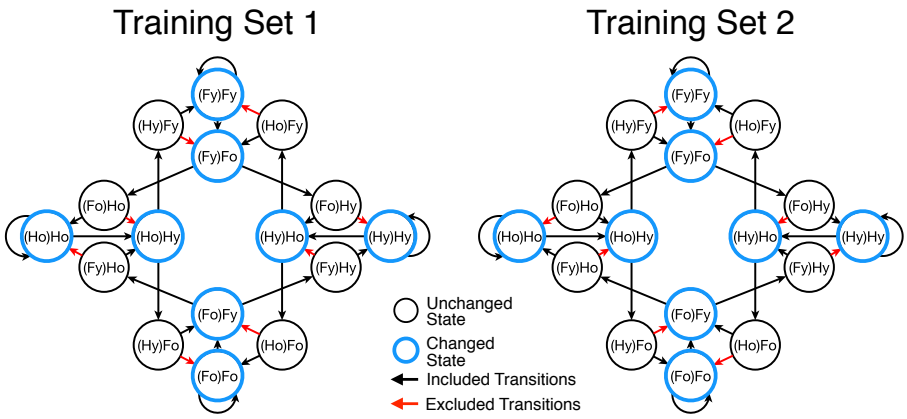

Figure S3. Transition exclusion analysis. In two separate sets of classifiers, we excluded one of the transitions leading to states with several incoming transitions (red arrows). The resulting classifiers therefore differed in the training set. Importantly, although each classifier was trained on data pertaining to all *states*, the data used to estimate the neural patterns evoked by the states marked in blue were based on different subsets of data, compared to the standard classifier. For each training set, the frequency of excluded transitions being decoded from the rest scans was compared to a matched set of included transitions (included transitions to the same blue states). Results were averaged across both training sets.

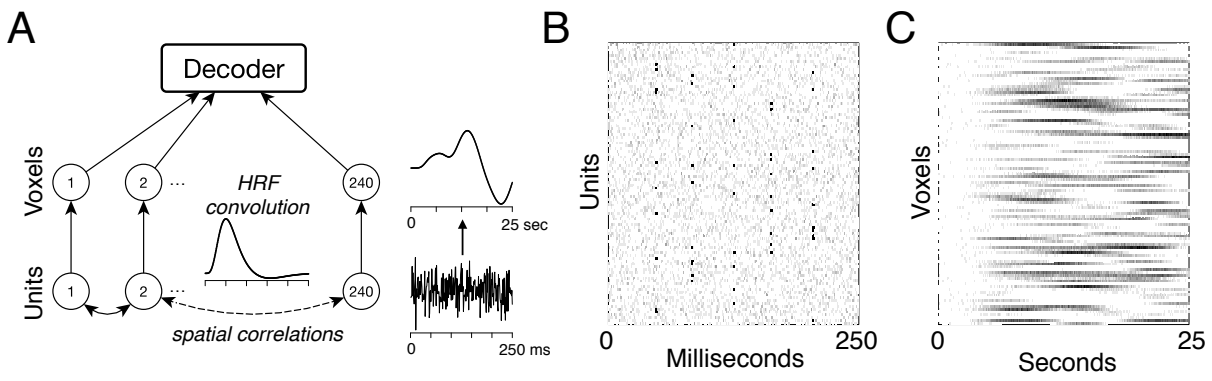

Figure S4. Sequence simulation. *(A)*: Illustration of replay simulation procedure. The simulation involved rapid activations in abstract 'units', which were convolved with a canonical HRF into BOLD-like signals, see text. *(B)*: One example of a sequence of pattern activations. *(C)*: Corresponding example of the pattern of simulated BOLD signals resulting from the sequence of pattern activations shown in B. Note the difference in time scale in panel C compared to panel B.

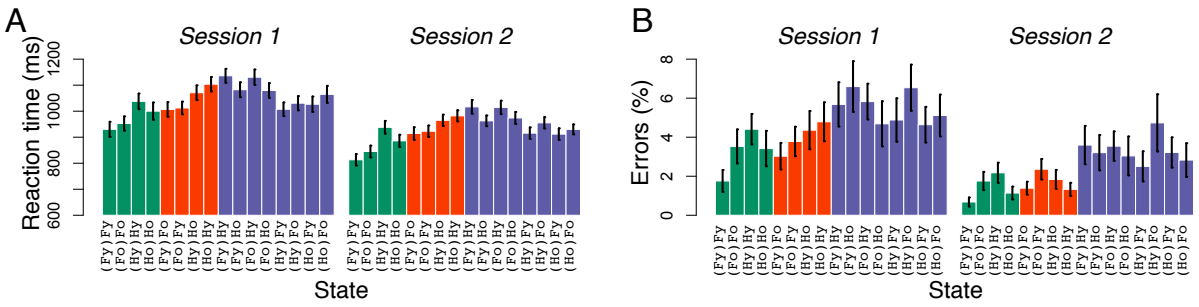

Figure S5. Effects of sessions and state identity on behavior. Reaction times *(A)* and error rates *(B)* are shown as a function of state identity (see x label) and session. Colors indicate the state class as in the main text. Bars represent standard errors of the mean.

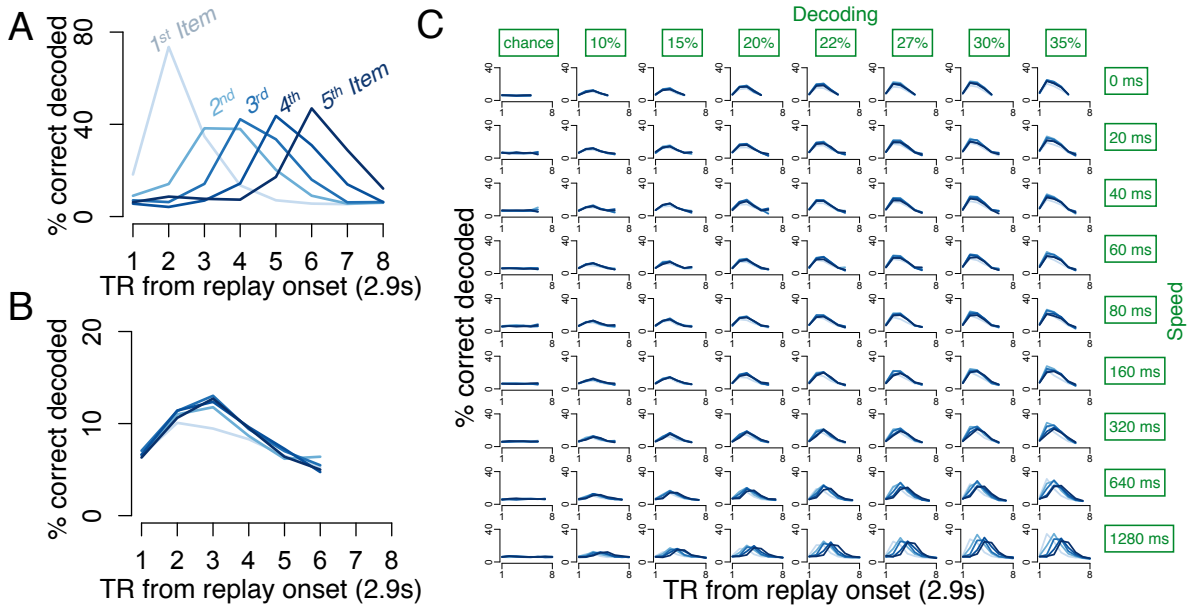

Figure S6. State decoding following replay events. All plots show the percentage of volumes (TRs) in which the respective state was correctly decoded, as a function of TR following the replay event (x-axis) and ordinal position of the state within the replayed sequence (color-coded lines, increasing from light blue to dark blue, see panel A). *(A)*: State identity decoding in simulations with slow replay (2560 ms between replayed states) and high decoding accuracy ($\sim$50%). *(B)*: Same figure as in (A), for a simulation with fast replay (40 ms between states) and low decoding accuracy ($\sim$ 12%). *(C)*: State identity decoding across all ranges of simulated replay speeds (see right side) and decoding accuracy (see top). Format as in panels A and B.

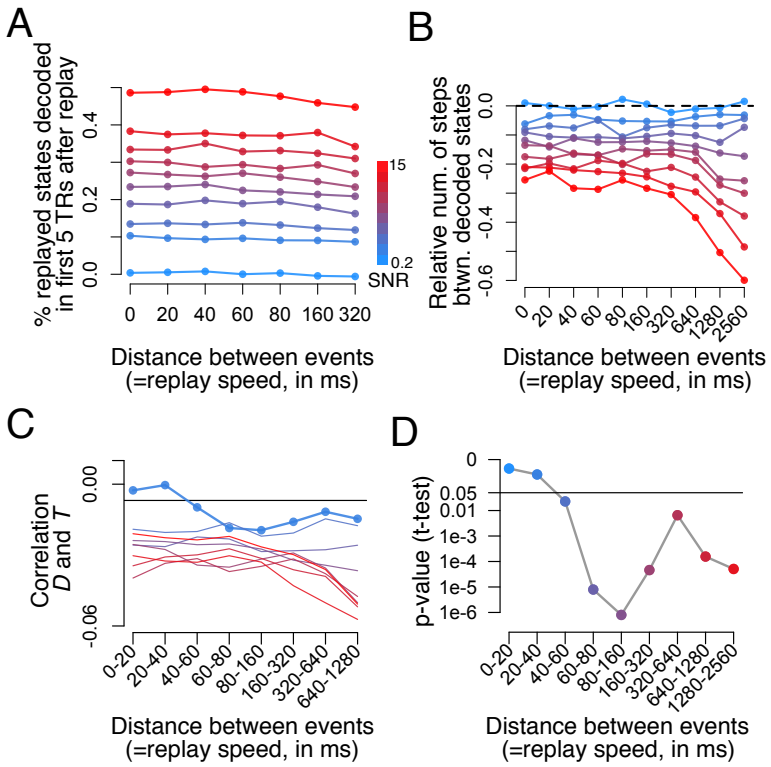

Figure S7. Sequence simulation. *(A)*: The percentage of states decoded in the first 5 TRs following a replay event that were part of the replayed sequence, for different levels of SNR/decoding accuracy (line color, see legend) and for different replay speeds (x-axis). *(B)*: The number of steps between successive states decoded during replay simulations as a function of SNR and replay speed (cf panel A). Plotted is the number of steps relative to the number of average steps found in a control simulation, in which random sequences of states were replayed. Cf. Fig. 3C, main text. *(C)*: The correlation between D and T, as in the analyses reported in the main text. The figure highlights this sequenceness measure as a function of replay speed for simulations with matched decoding accuracy (i.e., at 12% decoding, bold line). For comparison, the thin lines show simulations with higher SNR. The horizontal line shows the average correlation obtained in a permutation test. *(D)*: T tests comparing the correlations shown in (C) between replay and random sequence events (black horizontal line in C). The figure shows the resulting p values for the simulation with matched SNR, as a function of replay speed. The horizontal line represents the conventional $p < .05$ threshold.

Condition	Rand.	Seq.	Cond × Seq	AIC
PRE	23.16	–	–	3651.2
PRE	22.48	0.84	–	3648.6
TASK$_{pre}$	16.59	–	–	3661.5
TASK$_{pre}$	14.86	1.10	–	3656.9*
PRE & TASK$_{pre}$	25.44	0.75	–	7238.9
PRE & TASK$_{pre}$	28.66	0.41	0.67	7233.7*
INSTR	38.65	–	–	10044
INSTR	36.93	0.80	–	10044
TASK$_{instr}$	38.59	–	–	10166
TASK$_{instr}$	36.04	0.91	–	10155*
PRE & TASK$_{instr}$	50.17	0.86	–	19997
PRE & TASK$_{instr}$	51.20	0.78	0.10	19998

Table S1. Mixed effects modeling results. The structure of the models is the same as in the main text, see Methods. Here we used the long term average number of steps between states instead of the minimum number of steps between two states as a distance metric. Otherwise the models are identical to the models reported in the main text. The first column indicates from which condition(s) the data came. TASK$_{pre}$ refers to data from the task condition that was matched in length to the PRE condition, to equate power. TASK$_{instr}$ refers to TASK data matched to the INSTR condition. Columns 2-4 indicate fixed effects estimates derived from the models. AIC: Aikaike information criterion. *: significant difference in a log likelihood ratio test comparing the indicated model to the model in the previous line (performed only for nested models).

Condition	Rand.	Seq.	Cond × Seq	AIC
PRE	11.51	–	–	3583.6
PRE	11.92	0.05	–	3584.1
TASK$_{pre}$	4.17	–	–	3431.4
TASK$_{pre}$	3.07	-0.09	–	3431.8
PRE & TASK$_{pre}$	13.56	-0.01	–	6925.6
PRE & TASK$_{pre}$	15.15	0.06	-0.129	6925.4
INSTR	20.11	–	–	9912.0
INSTR	19.29	-0.10	–	9905.7*
TASK$_{instr}$	15.96	–	–	9817.6
TASK$_{instr}$	14.98	-0.10	–	9811.6*
PRE & TASK$_{instr}$	25.37	-0.08	–	19505
PRE & TASK$_{instr}$	26.27	-0.08	-0.001	19507

Table S2: Mixed effects modeling results. The table has the same structure as above (Table S1). Importantly, here predicted sequences of states derived from data from the orbitofrontal cortex were used.