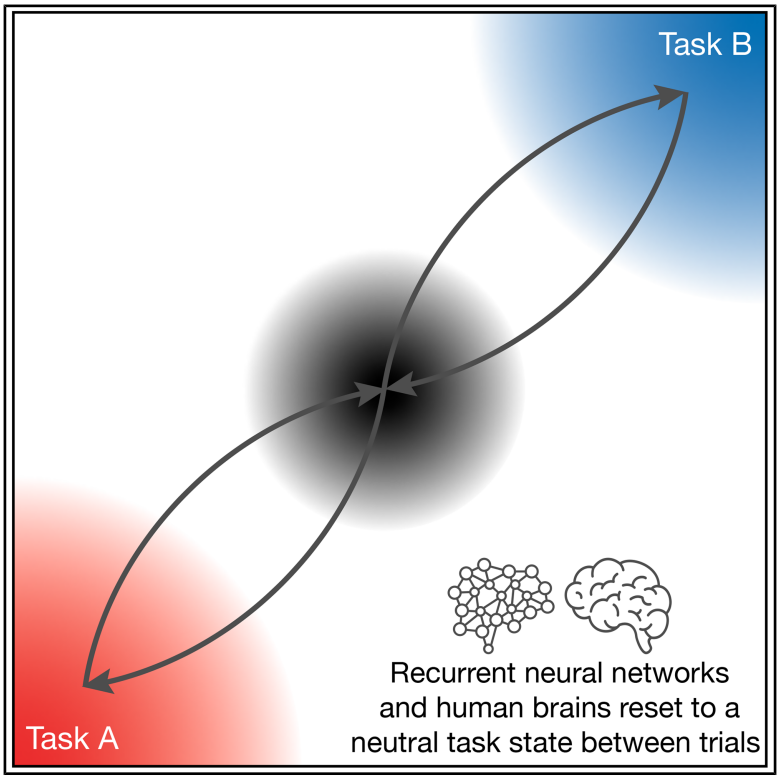
Article

Current Biology

Inter-trial convergence of neural task states supports cognitive flexibility

Graphical abstract



Highlights

- We formalized task-switching theories by training neural networks to switch tasks
- We modeled the state-space dynamics of neural networks and human brain recordings
- Neural networks and human brains reset to a "neutral" state before the next trial
- Control-dependent task resetting unifies classic theories of cognitive flexibility

Authors

Harrison Ritz, Aditi Jha,
Nathaniel D. Daw, Jonathan D. Cohen

Correspondence

hjr@queensu.ca

In brief

Ritz et al. show that RNN models of task switching reconcile long-standing theories of how people prepare for upcoming tasks. Using latent state-space models to compare preparatory dynamics in RNNs and human brain recordings, they find that effective performance depends on returning to a neutral task state between trials.

Ritz et al., 2026, Current Biology *36*, 1–13
September 7, 2026
https://doi.org/10.1016/j.cub.2026.07.028

CellPress

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

Current Biology

CellPress
OPEN ACCESS

Article

Inter-trial convergence of neural task states supports cognitive flexibility

Harrison Ritz,[1,2,6,*] Aditi Jha,[1,3,4] Nathaniel D. Daw,[1,5] and Jonathan D. Cohen[1,5]
[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[2]Centre for Neuroscience Studies, Queen's University, Kingston, ON K7L 3N6, Canada
[3]Department of Statistics, Stanford University, Stanford, CA 94305, USA
[4]Wu Tsai Neuroscience Institute, Stanford University, Stanford, CA 94305, USA
[5]Department of Psychology, Princeton University, Princeton, NJ 08544, USA
[6]Lead contact
*Correspondence: hjr@queensu.ca
https://doi.org/10.1016/j.cub.2026.07.028

SUMMARY

The ability to switch between tasks is a core component of human intelligence, yet a mechanistic understanding of this capacity has remained elusive. Long-standing debates over how task switching is influenced by preparation for upcoming tasks or interference from previous tasks have been difficult to resolve without quantitative neural predictions. We advance this debate by using state-space modeling to directly compare the latent task dynamics in task-optimized recurrent neural networks and human electroencephalographic recordings. Over the inter-trial interval, both neural networks and brains reset into a neutral task state, an effective dynamic motif that reconciles the roles of preparation and interference in task switching. These findings provide a quantitative account of cognitive flexibility and a promising paradigm for bridging artificial and biological neural networks.

INTRODUCTION

Humans have a remarkable capacity to flexibly adapt how they perform tasks.[1–4] This flexibility reflects a core feature of goal-directed cognition[5] and is a strong indicator of cognitive changes both across the lifespan[6,7] and in mental health.[8,9] Despite the centrality of cognitive flexibility in our mental life, we still have a limited mechanistic understanding of how people rapidly configure task processing to adapt to their moment-to-moment goals.

There is broad agreement in the task-switching literature that two core factors influence the ability to switch rapidly between stimulus-response mappings, often termed "task sets" (Figure 1A).[10,11] The first factor is "task set reconfiguration": when a new task is cued, *actively* switching between task sets.[12,13] The second factor is "task set inertia": interference from the previous task set that *passively* decays over time.[14,15] Here, we define active processes as those executed in anticipation of the upcoming task, and passive processes as those that obligatorily depend on the previous task. Note that active processes may be learned rather than planned, for example, in the weights of an artificial neural network trained via backpropagation-through-time.[16]

There have been long-standing debates over the relative contributions of active reconfiguration and passive interference,[2,3] with behavioral and neuroimaging experiments finding support for either or both factors.[17–23] However, this theorizing at the level of "factors," rather than quantitative process models, has made it difficult to generate precise predictions that could help adjudicate between theories of task switching from brain imaging experiments.

The current work aims to help clarify this debate by anchoring the distinction between reconfiguration and inertia on the operation of recurrent neural network models (RNNs) trained to perform the relevant tasks. Previous work has found that RNNs offer promising avenues for formalizing the influence of reconfiguration and inertia on task switching.[7,24–26] In these models, reconfiguration often arises from cue-driven activity or gating, whereas the influence of prior state often arises from recurrent activity that causes task representations to persist across trials. However, previous models have often relied on hand-crafted mechanisms to implement different components of task switching, which may not map well onto distributed computations in the brain. Moreover, there have been limited attempts to directly link the rich representational dynamics in artificial neural networks to recordings of brain activity during task switching.[27]

In the work reported here, we bridged the gap between dynamical theories of task switching and their neural mechanisms by leveraging two key innovations (Figure 1B). First, we generated quantitative neural predictions for how active reconfiguration and passive interference influence task switching by training task-optimized RNNs under theory-informed curricula. Second, we compared the task control signatures from these RNN models against two empirical human electroencephalography (EEG) datasets by fitting latent dynamical systems to both simulations and data. Our findings identify an effective dynamic motif in RNNs. RNNs that were trained to switch between tasks converged toward a "neutral" task state

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

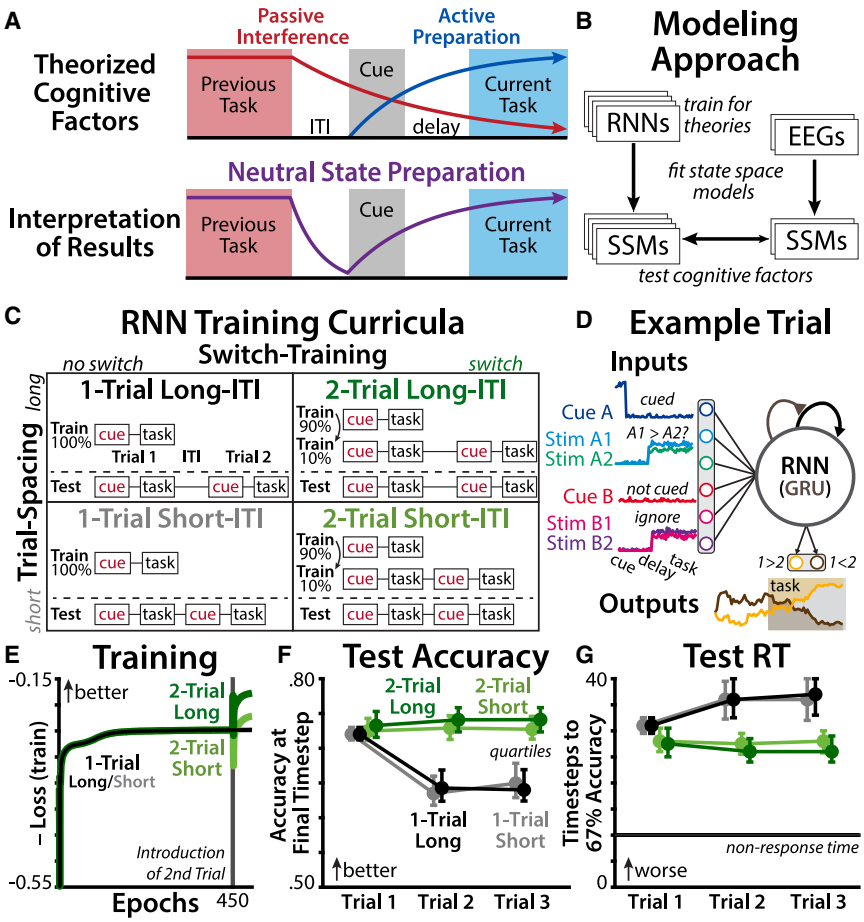


Figure 1. RNN training and task performance

(A) Top: the "passive interference" factor reflects the passive decay of previous task representations. The "active preparation" factor reflects preparation for upcoming tasks at the time of the cue. Bottom: this experiment finds evidence for the "neural state preparation" motif that revises both of these factors.

(B) The modeling approach for this paper was to (1) train RNNs to capture factors from different cognitive theories, (2) quantify the dynamic signatures of preparation in RNNs using state-space modeling and control theoretic analyses, and (3) compare the dynamic signatures in theory-trained RNNs against two EEG datasets.

(C) The *switch-training* curriculum (columns) manipulated RNNs' experience with task-switching, capturing preparatory strategies. The *trial-spacing* curriculum (rows) manipulated the ITI, capturing the influence of the previous task.

(D) Each trial consisted of a task cue (1-hot vector), a delay period (no inputs), and a target stimulus (two inputs per task). RNNs were trained to report which of the two task-relevant inputs had higher magnitude.

(E) Training loss over training epochs. All networks had the same single-trial training for the first 450 epochs, and then 2-Trial networks were trained on pairs of trials. Input noise was included on training trials.

(F and G) Mean final-timestep accuracy (F) and median timesteps to 67% accuracy (G) during a test phase (novel three-trial sequences, without input noise). Error bars reflect quartiles of the RNN distribution.

See Figure S1 for switch cost analysis.

during the inter-trial interval (ITI), unlike those trained to perform individual tasks. RNNs trained to switch under short ITIs had insufficient time to reach this neutral state, resulting in interference from the previous task. Critically, the neural signatures corresponding to this dynamic motif were closely mirrored in the EEG datasets, reconciling the influences of reconfiguration and interference.

RESULTS

Task-optimized RNNs learn to switch tasks

We used RNNs with gated recurrent units (GRUs)[28] to model how people flexibly switch between tasks (STAR Methods). These models have both the recurrent dynamics and gating mechanisms that have been proposed as important components of task switching.[24–26] Critically, these models provide us with full access to the latent computations. To isolate the effects of reconfiguration and prior tasks, we trained identical RNNs under the influence of two different task factors: "switch training" and "trial spacing" (Figure 1C).

The switch-training factor examined how RNNs actively prepared for upcoming tasks by varying their experience with switching. "1-Trial" RNNs were trained to perform only one trial per sequence, which did not provide experience with task switching during learning. In each sequence, they received a one-hot task cue, a delay period, and then, during a "task" period, they had to compare the relative magnitude of two task-relevant inputs (four simultaneous inputs, two per task; Figure 1D).[29,30] RNN states were read out through a linear transformation and logistic nonlinearity, and we trained RNNs to produce outputs that matched the correct response. We included a high level of input noise (cue signal-to-noise ratio [SNR]: 0 dB; trial SNR: −16.5 dB) to make the task difficult enough that it encouraged RNNs to actively prepare for upcoming tasks and to learn robust processing dynamics.

Whereas 1-Trial RNNs only ever performed one trial per sequence, "2-Trial" RNNs were also trained to switch and repeat tasks (Figure 1C, right column). 2-Trial RNNs performed single-trial sequences for the first 90% of their training and were then trained with two trials per sequence for the remaining 10% (equal probability of switching or repeating tasks). This addition of task switching late in training was designed to allow the RNNs to first develop a high degree of competence in context-dependent decision-making before learning how to transition between tasks, inspired by classic theories of how control intervenes when tasks are over-learned.[31,32] This training procedure provided a theory-informed comparison between 1-Trial RNNs trained only on decision-making (as a test of inertia) and 2-Trial RNNs trained to prepare for upcoming tasks (as a test of reconfiguration).

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

The trial-spacing factor examined the influence of prior task states by varying the ITI. This factor replicates classic behavioral experiments designed to test the inertia theory,[17] capturing how task switching is influenced by the temporal proximity to the previous trial. We manipulated temporal proximity by training RNNs with either short ITIs (20 timesteps; Figure 1C) or long ITIs (60 timesteps).

After training these four sets of RNNs ($N = 512$ per curriculum; Figure 1E), we tested their generalization on novel three-trial sequences, always with the same ITI durations as during training. All RNNs had fairly similar first-trial performance, suggesting a similar capacity for basic task processing (Figures 1F and 1G). However, 2-Trial RNNs had much better performance on the second trial of the sequence, especially when switching tasks (Figure S1). Despite their switch training, 2-Trial RNNs still exhibited poorer performance at shorter ITIs (trial 2 test loss: $d = 0.32$; trial 3 test loss: $d = 0.36$), consistent with a persistent influence of previous task states. On the third trial of the sequence, novel for all RNNs, 2-Trial RNNs maintained their performance advantage over 1-Trial RNNs, consistent with learning a generalizable reconfiguration strategy.

Quantifying latent dynamics with state-space models

We next turned to characterizing how RNNs' latent task dynamics depended on switching training and trial spacing. One common method for quantifying the dynamics of RNNs is to find locally linear representations around the model's fixed points (or "slow points").[33] However, comprehensively exploring fixed points in input-driven networks can be fraught, and this procedure does not allow for direct comparisons to neural recordings. Instead, we used a high-dimensional latent variable model to approximately linearize the entire RNN trajectory, as explained below.

We fit a linear-Gaussian state-space model (SSM; also commonly called a "latent linear dynamical system") to each RNN's hidden-unit activity.[34,35] SSMs have long been used in computational neuroscience to characterize latent dynamics in multiunit neural recordings.[36–38] Despite the nonlinear activation functions in real and artificial neurons,[39] linear models are often highly predictive of brain activity across measurement scales.[34,40,41] SSMs have several advantages over traditional dimensionality reduction techniques (e.g., principal-component analysis [PCA]) and sensor-level encoding analyses. These benefits include providing a generative model of the neural dynamics[42]; approximately linearizing the underlying dynamical system, which can be aided by projecting the observations into a higher-dimensional latent space[43]; and robustness to temporal jitter[44] (for discussion, see STAR Methods). Critically, SSMs can be characterized using a suite of tools from dynamical systems theory and control theory,[45,46] providing an interpretable alternative to nonlinear methods like hidden Markov models[47–49] and RNNs.[50,51]

The SSM generative model consists of a set of latent factors (state vector $\boldsymbol{x}$; Figure 2A; STAR Methods) that evolve through recurrence (A), inputs (with input vector $\boldsymbol{u}$ and input matrix B), and Gaussian "process noise" ($\boldsymbol{w}_t \sim \mathcal{N}(0, W)$). The observed RNN hidden states ($\boldsymbol{y}$) were "read out" from the latent factors according to an observation matrix (C) and Gaussian "observation noise" ($\boldsymbol{v}_t \sim \mathcal{N}(0, V)$).

$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + B\boldsymbol{u}_{t-1} + \boldsymbol{w}_t, \quad \text{(Equation 1)}$$

$$\boldsymbol{x}_{t=1} = B_0\boldsymbol{u}_0 + \boldsymbol{w}_1$$

$$\boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{v}_t$$

We captured task-switching dynamics by including both the previous and current task identities in the SSM inputs (see STAR Methods for the complete predictor list). To provide the model with additional flexibility, we expanded the inputs with a temporal spline basis (Figure S2). To capture the influence of the previous trial's task state, the previous task identity was included in the predictors for the initial condition ($B_0^{\text{prevTask}}\boldsymbol{u}_0^{\text{prevTask}}$; STAR Methods).

We reduced the dimensionality of the RNN hidden states using PCA, retaining components that accounted for 99% of the variance (49–76 PCs). We fit the SSM to an epoch encompassing the cue period, delay period, and trial onset. We developed an open-source analysis package for estimating the maximum *a posteriori* SSM parameters through a combination of subspace identification and expectation maximization[52–54] (STAR Methods).

We found that the best cross-validated fit came from high-dimensional ("over-determined") SSMs, which had more factors (112) than observation dimensions (49–76 PCs) or hidden units (108; Figure 2B). This "lifting" of the dimensionality can allow a linear model to better approximate an underlying nonlinear system.[43,55] Consistent with this principle, the best-fitting model had high predictive accuracy for future timesteps on held-out trials (Figure 2C) and had better cross-validated accuracy than autoregressive encoding models without latent states (R^2 vs. null: 95% confidence interval [CI] [.16,.30]; STAR Methods). We further validated the estimation procedure using parameter recovery, generating a synthetic dataset from a fitted SSM and then re-fitting the model to these data. We found good recovery of the ground-truth parameters (Figure 2D; STAR Methods), supporting the interpretability of this modeling approach.

SSMs capture latent dynamics in human EEG

Having validated our SSM fitting procedure on RNNs, we next applied this procedure to two previously reported human EEG datasets collected during complementary task-switching experiments. The first dataset was from an experiment by Hall-McMaster et al.,[56] in which they had 30 participants perform a cued task-switching paradigm while recording 61-channel EEG (Figure 3A; STAR Methods). On each trial, participants were cued whether to perform a "shape" or "color" task. Following a delay period, participants saw a colored shape and responded with a button press based on the cued dimension. The second dataset was from Arnau et al.,[57] in which they had 26 participants perform a similar task-switching paradigm while they recorded 128-channel EEG, here switching between color and orientation targets in a visual search task (Figure 3A; STAR Methods).

While the two EEG datasets differed in several respects, such as their tasks, countries of data collection, and electrode count, of primary interest here is that they had substantially different ITIs. The "Hall-McMaster" dataset had a much longer ITI (2,400–2,800 ms) than the "Arnau" dataset (800–1,000 ms).

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

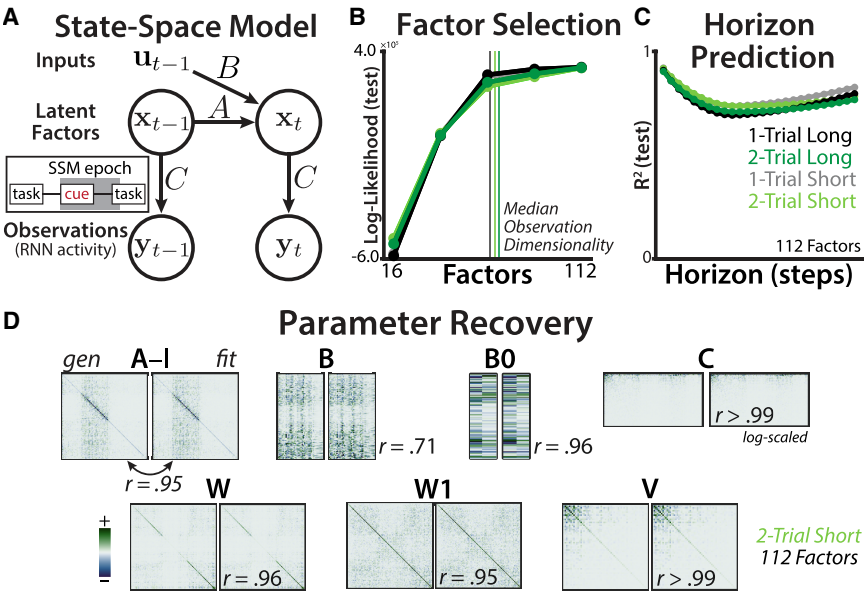


Figure 2. State-space analysis of RNN models

(A) Schematic of the SSM. Observations (***y***) are generated from latent factors (***x***), which evolve according to their intrinsic dynamics and inputs (***u***). Inset: SSM epoch included the cue period, the delay period, and the onset of the task.

(B) SSM log-likelihood in held-out trials at different numbers of factors. Log-likelihood is mean-centered within each network to show relative differences. Similar fits across curricula (refer to legend in panel C).

(C) SSM R^2 in held-out trials, at different prediction horizons. Similarly high predictive accuracy across curricula, even at a horizon of 25 timesteps (more than half the epoch).

(D) Ground truth (generating) and recovered (fit) parameters for an example RNN dataset. Parameters are linearly aligned to account for degeneracy in the latent system (STAR Methods). Colormaps are scaled separately for each parameter pair, symmetrically around zero.

See Figure S2 for information on input bases.

The ITI ratio between EEG experiments was similar to what we used to train long-ITI and short-ITI RNNs (approximately 3 to 1).

We fit SSMs to each EEG dataset, using the same fitting procedure that we used for RNNs. As with RNNs, we again found that the best-fitting models had more latent factors than EEG components (Hall-McMaster: 112 factors fit best, 14–27 PCs; Arnau: 128 factors fit best, 15–35 PCs; 112 factors used throughout; Figure 3B). SSMs provided better predictions of held-out EEG activity than both simpler autoregressive encoding models and nonlinear RNNs trained to predict EEG (Figure 3B; STAR Methods).

Visualizing the inferred EEG task trajectories, we found that task states differentiated over the course of the preparatory period (Figure 3C). These EEG task states appeared to support good performance: when EEG recordings were better aligned with the correct task state during preparation, participants tended to have faster reaction times (linear mixed model; Hall-McMaster: d = 0.59, p = 0.0073; Arnau: d = 0.75, p = 0.012; STAR Methods).

As with RNNs, over-determined SSMs made accurate predictions on held-out EEG trials for several timesteps into the future (Figure 3D). We further evaluated how well SSMs could capture the spectral profiles of the EEG data (Figure 3E). We generated 100 synthetic datasets from feedforward rollouts from our fitted SSMs and compared the power spectrum in these synthetic datasets against the empirical datasets. Across participants, SSMs accurately captured empirical power spectra (Hall-McMaster: R^2_{CV} 95% CI [0.97, 0.98], noise-normalized R^2_{CV} 95% CI [.99,.99]; Arnau: R^2_{CV} 95% CI [0.99, 0.99], noise-normalized R^2_{CV} 95% CI [0.99, 1.0]). As with RNNs, SSMs fitted to EEG datasets had excellent parameter recovery (Figure S3).

Active reconfiguration of neural task states

Having validated the predictive power of state-space modeling in RNNs and EEG, we next characterized how switch training and trial spacing altered RNNs' latent dynamics. We visualized the RNN hidden-unit activations in three dimensions by projecting the condition-averaged response onto a hidden-unit embedding estimated through singular value decomposition (SVD; Figures 4A–4D; STAR Methods).

We found that all RNNs exhibited similar first-trial dynamics, but their second-trial dynamics were qualitatively different across curricula. These visualizations revealed three dynamic signatures of RNNs' task-switching strategies, which we quantified using the fitted SSMs. Across all three signatures, we found that 2-Trial RNNs were a better match to human EEG than 1-Trial RNNs, consistent with active reconfiguration. Complementing these process-specific signatures, 2-Trial RNNs also had higher global similarity with the EEG trajectories (Figure S4). Moreover, differences between RNN datasets with long vs. short ITIs were recapitulated in the differences between EEG datasets with long vs. short ITIs, consistent with an influence of the previous task state.

Task energy

The first dynamic signature corresponded to the magnitude of task-dependent dynamics following a cue. Following the task cue, 2-Trial RNNs appeared to have much more robust dynamics than 1-Trial RNNs, especially when switching between tasks (Figure 4E; "task energy"). The deployment of reconfiguration processes on switch trials, compared with repeat trials, has long been a central focus of task-switching research.[21]

To isolate the dynamics that were attributable to task processing in our fitted SSMs, we estimated the "task state" ($\boldsymbol{x}^{\text{task}}$) as the latent trajectory that results from task inputs alone (with $B^{\text{task}}\boldsymbol{u}^{\text{task}}$ indicating task-specific inputs; STAR Methods):

$$\boldsymbol{x}_t^{\text{task}} = A\boldsymbol{x}_{t-1}^{\text{task}} + B^{\text{task}}\boldsymbol{u}_{t-1}^{\text{task}}, \qquad \text{(Equation 2)}$$

$$\boldsymbol{x}_{t=1}^{\text{task}} = B_0^{\text{prevTask}}\boldsymbol{u}_0^{\text{prevTask}}$$

To quantify the extent to which RNN trajectories were task-dependent, we used "Lyapunov analyses," a standard tool from control theory for quantifying the influence of control

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

CellPress
OPEN ACCESS

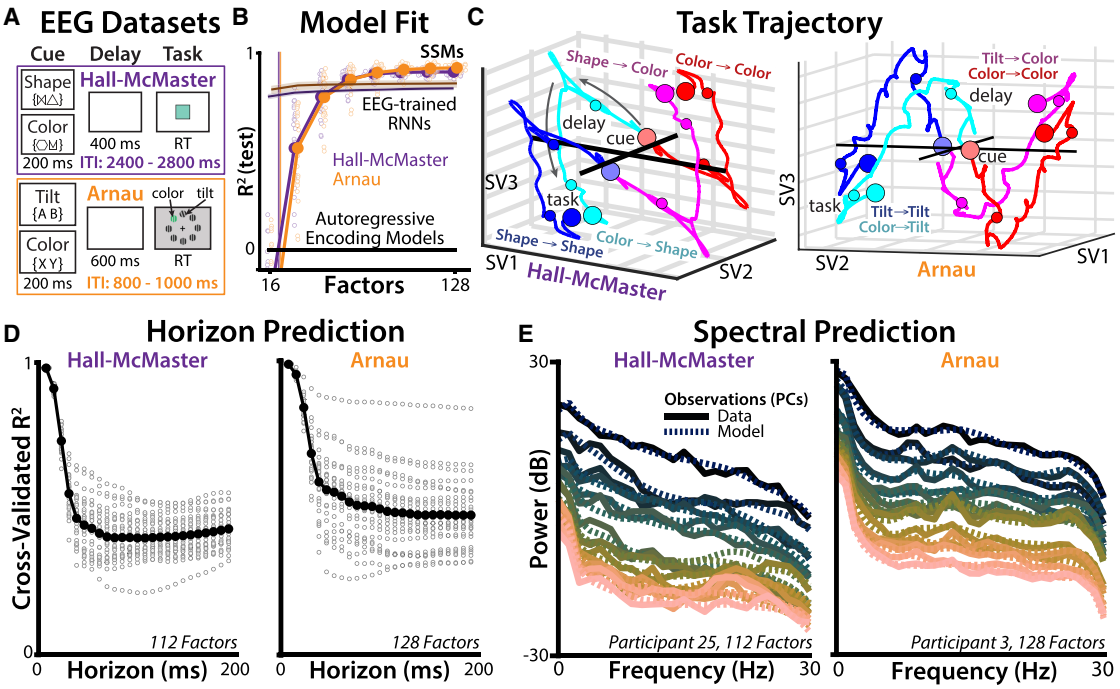


Figure 3. State-space analysis of EEG datasets

(A) Task protocol for the "Hall-McMaster" (purple) and "Arnau" (orange) datasets.

(B) Cross-validated R^2 at different numbers of latent factors. Zero indicates fit of autoregressive encoding model. Darkly colored lines indicate EEG-trained RNN models. Vertical lines indicate median observation dimensionality.

(C) Embedded SSM task trajectories for each task transition (SV: singular vector). Task is contrast-coded, so trajectories are symmetrical within switch condition. Left: Hall-McMaster, right: Arnau.

(D) Predictive accuracy of the model when forecasting at different timesteps into the future. Gray dots indicate individual participants.

(E) Similarity in the power spectrum for test-set observations (solid lines) and model simulations (dashed lines), shown for example participants. Power spectra are plotted for each observation dimension (i.e., principal component), aggregated across stimulated epochs.

See Figure S3 for EEG parameter recovery.

inputs.[45] Previous work has used similar methods for asymptotic "controllability analyses" of how well control inputs could drive a neural system with a given connectivity profile.[46,58,59] Importantly, these methods account for both the instantaneous input strength and how inputs propagate through recurrent dynamics. Here, we quantified the realized task energy using a recursive Lyapunov equation[60] (STAR Methods):

$$\mathcal{G}_t^{\text{task}} = A\mathcal{G}_{t-1}^{\text{task}}A^\top + \left(B^{\text{task}}\boldsymbol{u}_{t-1}^{\text{task}}\right)\left(B^{\text{task}}\boldsymbol{u}_{t-1}^{\text{task}}\right)^\top, \qquad \text{(Equation 3)}$$

$$\mathcal{G}_{t=1}^{\text{task}} = \left(B_0^{prevTask}\boldsymbol{u}_0^{\text{prevTask}}\right)\left(B_0^{prevTask}\boldsymbol{u}_0^{\text{prevTask}}\right)^\top$$

The task Gramian ($\mathcal{G}_t^{\text{task}}$) is analogous to the state covariance matrix attributable to task inputs, revealing how strongly inputs drive the latent state. We used the average task energy metric ($\text{trace}\left(\mathcal{G}_t^{\text{task}}\right)$) to compare the strength of task inputs on switch vs. repeat trials.

Comparing task energy between switch and repeat trials revealed a strong influence of switch training (Figure 5A). 1-Trial RNNs exhibited weak or absent differentiation between switch and repeat conditions, especially at short ITIs. By contrast, 2-Trial RNNs in both ITI conditions showed stronger task energy on switch trials than on repeat trials. Through their training on task switching, 2-Trial RNNs learned dynamics that were responsive to task demands.

In the EEG datasets, we found stronger task energy on switch trials than on repeat trials (Figure 5B). This profile of task energy was more consistent with 2-Trial RNNs than 1-Trial RNNs (contrasting 2-Trial vs. 1-Trial cosine similarity with EEG: Hall-McMaster: 95% CI [0.87, 1.2]; Arnau: 95% CI [1.3, 1.6]; Figures 5B and 5C). We confirmed that there was similar correspondence in RNNs for which the cue onset was made less predictable through variable ITIs (Figure S5).

Switch similarity

Complementing the observation that switch and repeat conditions differed in the strength of task encoding, the second dynamic signature characterized the alignment between switch and repeat trajectories (Figure 4E; "switch similarity"). Switch and repeat trajectories may differ in their velocity, but do they take a similar path? For 1-Trial RNNs, the trajectories on switch trials appeared to be quite different, even opposite, from the trajectory on repeat trials. By contrast, 2-Trial RNNs' trajectories on switch appeared to closely mirror the trajectories on repeat trials. Starting from a neutral state between the two task states, 2-Trial RNNs appeared to reproduce a similar trajectory on the second (switch or repeat) trial as they did on the first trial.

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

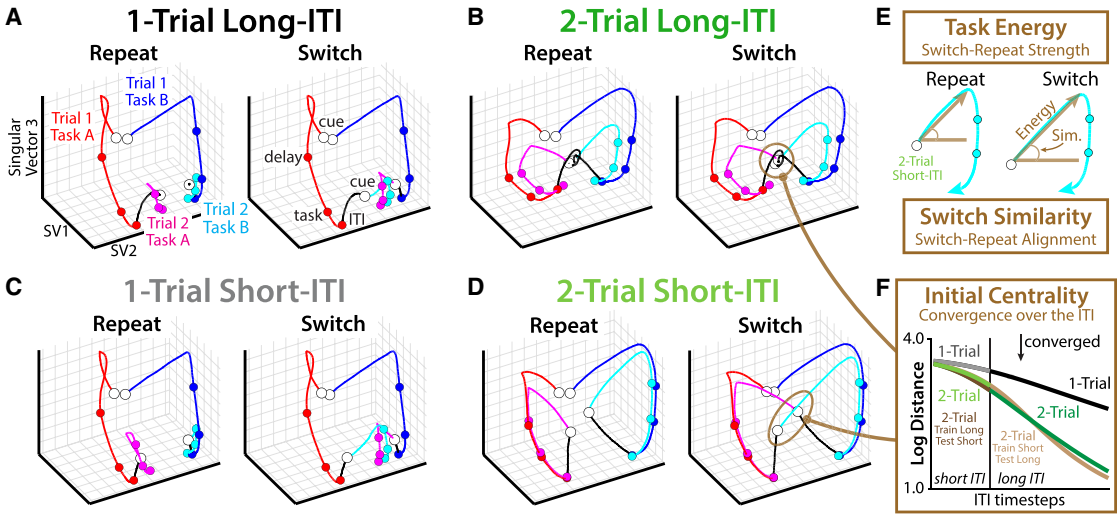


Figure 4. Dynamic signatures of task preparation in RNN hidden-unit activity

(A–D) Low-dimensional embedding of example RNN hidden-unit activity for different task transitions. Plotted for (A) 1-Trial Long-ITI, (B) 2-Trial Long-ITI, (C) 1-Trial Short-ITI, and (D) 2-Trial Short-ITI. White dots indicate cue onset, colored lines indicate trial type, and a black line indicates ITI.

(E) Comparing switch and repeat trajectories based on their displacement (*task energy*; top) and alignment (*switch similarity*; bottom). In 2-Trial RNNs, switch trials appeared to have more robust dynamics (task energy) along a similar trajectory (switch similarity) as repeat trials. In 1-Trial RNNs, this pattern was largely absent.

(F) *Initial centrality*: convergence toward a neutral state during the ITI in 2-Trial RNNs, unlike 1-Trial RNNs. Inset: Euclidean distance between RNN hidden states following each task, over the course of the ITI.

See Figure S4 for global RNN-EEG comparison.

We compared the task trajectories across switch and repeat trials by computing the cosine similarity between their task states at different temporal lags[61,62] (Figure 6A; STAR Methods). 1-Trial RNNs had persistent negative alignment between switch and repeat trials, consistent with opposing dynamics along a "task" axis. By contrast, 2-Trial RNNs had positive alignment between switch and repeat trials, though this varied as a function of trial spacing. Whereas 2-Trial RNNs with long ITIs exhibited this positive similarity throughout the trial, this alignment appeared later in 2-Trial RNNs with short ITIs.

In the EEG datasets, we found aligned task encoding across switch and repeat trials (Figure 6B). As with task energy, this pattern in EEG was more consistent with 2-Trial RNNs than with 1-Trial RNNs (contrasting 2-Trial vs. 1-Trial cosine similarity with EEG: Hall-McMaster: 95% CI [0.18, 0.83], Arnau: 95% CI [0.076, 0.46]; Figure 6C; STAR Methods).

Initial centrality

The third dynamic signature was the convergence of task states toward a neutral state during the ITI. At the end of the first trial, RNNs were in distinct task states (e.g., Figure 4A; "ITI"). During the ITI, the distance between these states decreased as RNNs

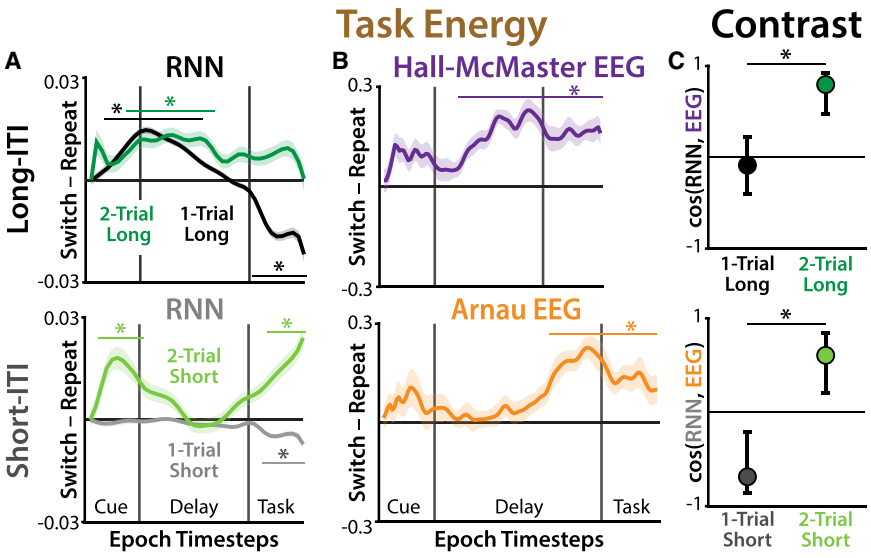


Figure 5. RNNs and humans show switch-elevated task encoding

(A and B) Difference in task energy between switch and repeat trials, for (A) RNNs and (B) EEG. Asterisks indicate $p < 0.05$ (threshold-free cluster enhancement [TFCE]-corrected).

(C) Bootstrapped cosine similarity between EEG and 1-Trial vs. 2-Trial RNNs. Asterisks indicate $p < 0.05$ (bootstrap test).

See Figure S5 for variable-ITI control analysis.

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

CellPress
OPEN ACCESS

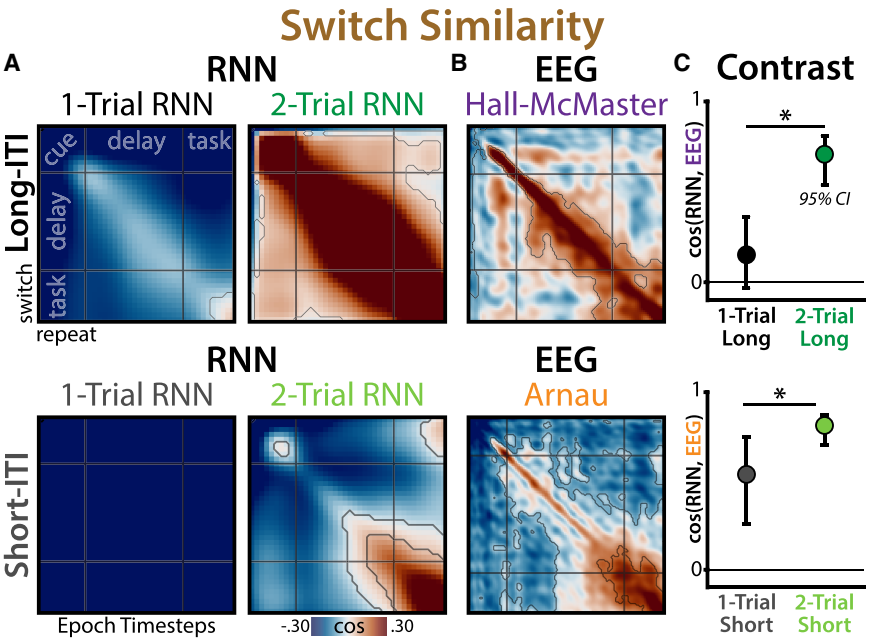


Figure 6. RNNs and humans show ITI-dependent profiles of representational geometry

(A and B) Cross-lagged similarity between SSM task states on switch vs. repeat trials, for (A) RNNs and (B) EEG. Cardinal lines indicate boundaries between cue, delay, and task events. Contours indicate $p < 0.05$ (TFCE-corrected).

(C) Bootstrapped cosine similarity between EEG and 1-Trial vs. 2-Trial RNNs.

moved into a more similar state (Figure 4F; "initial centrality"). This convergence was faster for 2-Trial RNNs than for 1-Trial RNNs, despite starting the ITI at a similar initial distance. To verify that this convergence didn't reflect 2-Trial RNNs over-learning specific ITI conditions, we swapped the ITI condition at test (e.g., trained on long-ITI then tested on short-ITI). We found that RNNs with swapped ITI had a similar rate of convergence to those with matching ITI, consistent with RNNs learning generalizable dynamics (Figure 4F).

Starting the second trial at a common state may explain why 2-Trial RNNs had stronger switch similarity than 1-Trial RNNs: if RNNs started near the midpoint between the task states, this would produce symmetrical trajectories. It may also explain why the ITI duration influenced switch similarity in 2-Trial RNNs: shorter ITIs left less time to reach the "neutral state" between tasks. Returning to the midpoint between task states likely emerged from optimization because it is a sensible strategy to equate the initial distance between task states when the upcoming trial is unpredictable.

To understand *where* RNNs had converged by the end of the ITI, we used SSMs to quantify whether the initial conditions of the second trial were equidistant from the states for task A and task B. We measured the relative Euclidean distance of each task state to the initial conditions over the course of the epoch (STAR Methods):

$$\delta_t = \frac{\| \boldsymbol{x}_1^A - \boldsymbol{x}_t^A \|_2 - \| \boldsymbol{x}_1^A - \boldsymbol{x}_t^B \|_2}{\| \boldsymbol{x}_1^A - \boldsymbol{x}_t^A \|_2 + \| \boldsymbol{x}_1^A - \boldsymbol{x}_t^B \|_2,} \qquad \text{(Equation 4)}$$

A value of zero indicated that the initial conditions were equidistant from the two task states at that timestep, whereas a negative value indicated that the RNNs were closer to the task state from the previous trial. In 1-Trial RNNs, the initial state was far from the midpoint throughout the epoch, without strong differentiation across ITI conditions (Figure 7A). In 2-Trial RNNs, however, the initial state was near the midpoint of the task states during the task period, and RNNs with longer ITIs achieved a more neutral initial state, especially during the cue and delay periods. This confirms that RNNs transitioned into a neutral state over the ITI, well-positioned for the upcoming trial phases.

We performed several control analyses to validate that 2-Trial RNNs learn to converge into a task-neutral state. First, we confirmed that convergence into a task-neutral state does not occur when we include ITI response penalties (in the absence of switch training), indicating that the task-neutral state differed from the response-neutral state (Figures S6A–S6C). Second, we confirmed that RNNs were not merely converging toward a low-energy state like the origin. During the ITI, we found that the distance between ITI trajectories following each task shrunk much faster than their mean distance from the origin, again consistent with a task-relevant neutral state (Figure S6D). Third, we found that this convergence reflects attractor dynamics, with nonlinear contraction analysis revealing faster convergence in 2-Trial RNNs than 1-Trial RNNs, particularly for the fastest-converging subspace (Figures S6E–S6G). Finally, we found that this convergence depended on task expectations, with RNNs trained under a higher probability of repeating or switching tasks converging to an ITI state that was biased toward the previous or upcoming tasks, respectively (Figures S6H–S6J).

In the two EEG datasets, we found that the initial conditions were close to the midpoint of the task states (Figure 7B). This pattern closely matched 2-Trial RNNs but not 1-Trial RNNs (contrasting 2-Trial vs. 1-Trial R^2 with EEG: 95% CI [8.3, 13]; Figure 7C). As with 2-Trial RNNs, the initial conditions were closer to the midpoint in the EEG dataset with longer ITIs (two-sample t test on midpoint score between EEG experiments, averaged over cue period: $d = 1.16$, $p = 6.54 \times 10^{-5}$; averaged over delay period: $d = 0.75$, $p = 0.0068$).

We sought to directly test whether putative differences in the extent of ITI convergence between EEG experiments could also account for the differences in switch similarity (see Figure 6). If equating the degree of ITI convergence between datasets could recover the same pattern of switch similarity, this would support a shared explanation for these phenomena. In our SSMs, the initial centrality is most naturally captured by the strength of previous-task encoding at the initial conditions ("initial interference"). We found that Arnau had greater initial interference than Hall-McMaster (Figure 7D). When we re-

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

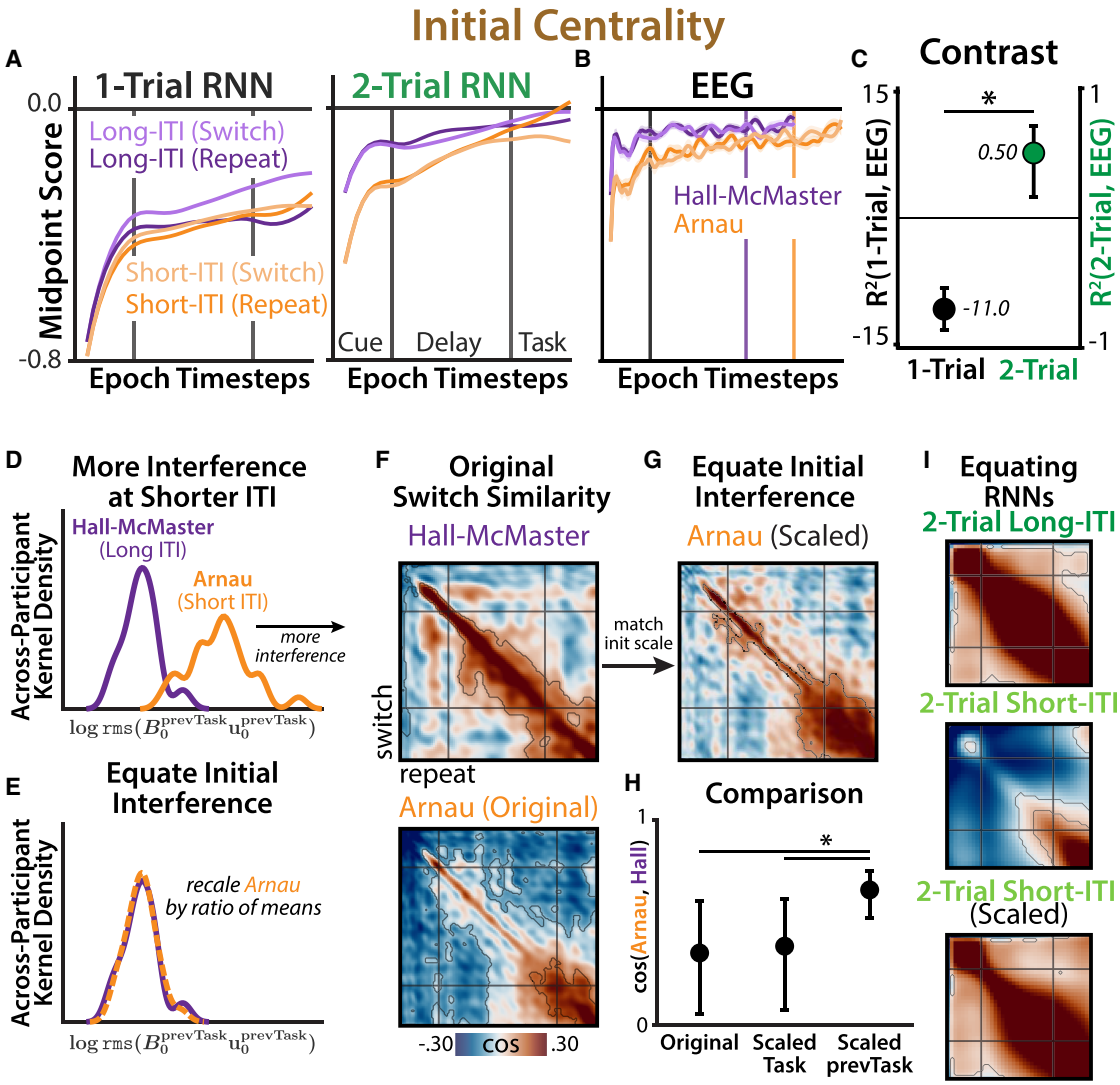


Figure 7. RNNs and humans show similar signatures of inter-trial convergence

(A and B) Relative distance between SSM initial conditions and each task state for (A) RNNs and (B) EEG. A value of zero indicates that the initial conditions are equidistant from the task states at that timestep. Negative values indicate a bias towards the previous task.

(C) Bootstrapped R^2 between EEG and 1-Trial vs. 2-Trial RNNs. Note that the *y* axes have different scales for better visualization. This R^2 metric can have negative values when the deviation exceeds the variance (STAR Methods). Asterisks indicate $p < 0.05$ (bootstrap test).

(D) Strength of previous-task encoding at the initial conditions.

(E) Matching initial interference between datasets.

(F) Cross-lagged similarity between SSM task states on switch vs. repeat trials for the two EEG datasets. Compare to Figure 6B.

(G) Switch similarity in Arnau after rescaling the initial interference to match Hall-McMaster.

(H) Similarity between datasets without rescaling ("Original"), rescaling the initial current-task encoding ("Scaled Task"), and rescaling the initial previous-task encoding ("Scaled prevTask"). Asterisks indicate $p < 0.05$ (bootstrap test).

(I) Analogous rescaling effects in 2-Trial RNNs.

See Figure S6 for convergence control analyses and Figure S7 for analyses of RNN gating mechanisms.

better reproduced Hall-McMaster's pattern of switch similarity (Figures 7F and 7G), unlike when we rescaled the initial encoding of the *upcoming* task (Figure 7H; see STAR Methods). Rescaling 2-Trial RNNs to match the initial interference in short-ITI and long-ITI networks produced a similar outcome (Figure 7I). This confirms that the partial progress toward a neutral task state is sufficient to account for some between-study differences in task trajectories.

We trained our RNNs with explicit gating mechanisms (GRUs) that provide improved long-range learning[63] and control over integration timescales.[64] We explored whether gating offers a plausible low-level mechanism for active reconfiguration, consistent with classic models of cognitive control.[65,66] We found that 2-Trial RNNs with ablated GRU gates, or with gates that were frozen after the initial single-trial learning, had impaired task-switching performance (Figures S7A–S7C;

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

STAR Methods). Moreover, we found that 2-Trial RNNs trained without any gates poorly reproduced the dynamic signatures observed across the two EEG datasets (Figures S7D and S7E; STAR Methods). While these results are promising, neural recordings at a finer spatial scale are needed to resolve these low-level mechanisms, especially given the parameter degeneracy in task-optimized neural networks.[67–69]

These findings suggest an important augmentation of the standard view of how neural systems flexibly switch between tasks. These results suggest that, rather than switching directly from one task state to another at the time of the cue, the brain exhibits task-relevant dynamics during the inter-trial period, dynamics that are consistent with RNNs that have learned to prepare for upcoming tasks. The brain moves toward an intermediate, "neutral" state before a task cue, a state that is roughly equidistant between the upcoming task states. When the task cue appears, RNNs and brains transition into the appropriate task state, with faster dynamics when they need to switch tasks. Together, these results demonstrate a striking correspondence between task-optimized RNNs and human neural dynamics, consistent with a common set of dynamics for flexible information processing.

DISCUSSION

Psychologists have debated the latent processes underlying cognitive flexibility[2,3] for decades, at least in part because a quantitative account of the underlying neural mechanisms has remained elusive. To explore such an account, we trained neural networks on curricula that manipulated the primary factors theorized to influence task-switching performance: active preparation for upcoming tasks and interference from previous tasks. To quantify how these factors affected RNN dynamics, we fit SSMs to the RNN hidden-unit activity. This state-space modeling approach allowed us to precisely compare task dynamics in RNNs and human brain recordings using a common statistical model.

These analyses revealed that RNNs trained to switch tasks exhibit dynamics in which they transition into a neutral task state during the ITI, like a tennis player recovering to the center of the court after a shot. This represents a novel augmentation to existing task-switching theories, reconciling the effects of previous task states with active reconfiguration. It takes time to reconfigure into the neutral state, which makes switch and repeat dynamics more symmetric at longer ITIs. Critically, these dynamics depended on RNNs being trained to switch between tasks, consistent with a learned control strategy.

We tested the extent to which this RNN model could explain human brain activity during task switching. We deployed the same inferential model and control theoretic analyses on two EEG datasets, which we then compared with the trained RNNs. These analyses revealed that human brain activity exhibits strikingly similar dynamics to those observed in RNNs that learn to switch between tasks. In particular, the human brain appears to replicate the same dynamic motif as RNNs, moving into a neutral state between trials. The convergence of this approach highlights the potential benefits of "mentally resetting" between difficult tasks.

The two EEG datasets diverged in several respects beyond ITI, such as their tasks, the country where the data were collected, and their electrode count. This array of divergences between datasets precludes a definitive conclusion about whether ITI-dependent task convergence produced the difference between EEG datasets. However, the RNN simulations demonstrate that ITI convergence is a sufficient and parsimonious explanation for these differences. Future experiments should explore whether ITI-dependent effects persist under better-controlled manipulations and whether other experimental manipulations produce the same complex patterns of neural dynamics that we found in RNNs.

These findings provide a valuable framework for understanding the neuro-computational mechanisms underlying effects that have been the target of classic psychological theories of task-switching. Our results are consistent with a role for active preparation[12]; however, our observations of preparation on repeat trials and before task cues extend these accounts. We also find evidence for previous-task interference[14]; however, our results are not consistent with interference being due to the passive decay of task representations as such (cf. 1-Trial RNNs). Rather, they appear to reflect movement to a neutral state of preparation before the upcoming task cue. While previous work has proposed that task switching reflects both of these factors,[17,25] our findings offer a unification of these factors in terms of dynamic transitions into a neutral state, grounded in the optimization of task performance by gated neural networks. This revised theory offers productive directions for future research. Future experiments should explore the circuit-level mechanisms that support the neutral state motif (Figure S7), how this motif depends on the probability of switching tasks (Figure S6H), and how the location of neutral task states changes under different task pairings.[70]

These theoretical insights were enabled by the use of linear-Gaussian SSMs to approximate nonlinear systems such as RNNs and human brains. In the over-determined regime (more latent than observed dimensions), SSMs exhibited a powerful combination of predictive power and interpretability. While this state-space approach has considerable promise for human neuroscience, some caveats remain to be addressed. First, the full expressivity of our dynamics came from modeling time-varying inputs using a descriptive model (spline basis) rather than a process model. Future research should develop process models that elaborate on the task-dependent inputs (e.g., using optimal control[46,58,71,72]). A second caveat is that switch-trained RNNs showed weaker switch costs than expected from human behavior (Figure S1), despite similar preparatory neural dynamics. This is consistent with theories that posit additional contributions to switch costs beyond preparation (e.g., associative learning),[11,18,73] and future research should extend these findings to better accommodate behavioral switch costs.[25,74] Third, we chose to train RNNs under a blocked curriculum rather than interleaving experience with single and double trials. Future work should explore how interleaved training changes preparatory dynamics in humans and RNNs. Despite these caveats, the combination of neural network modeling, latent inference, and control theory used in the current experiment has the potential to help scientists better understand the dynamics of distributed neural systems that enable flexible cognition, both *in vivo* and *in silico*.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to, and will be fulfilled by, the lead contact, Harrison Ritz (hjr@queensu.ca; https://harrisonritz.github.io).

Materials availability

This study did not generate any new, unique reagents.

Data and code availability

- EEG data have been deposited at the OSF: Open Science Framework by the original authors and are publicly available as of the date of publication at https://osf.io/kuzye/ ("Hall-McMaster" dataset) and https://osf.io/ndgst/ ("Arnau" dataset).
- All original code has been deposited at GitHub at https://github.com/harrisonritz/convergent-neural-dynamical-systems (analysis code) and https://doi.org/10.5281/zenodo.14511205 (state-space analysis package) and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We are extremely grateful to Sam Hall-McMaster and Stefan Arnau for making their datasets open access and for providing additional support. Thanks to Christopher Langdon for sharing PyTorch code for context-dependent decision-making tasks, Caroline Jahn for feedback on an early draft of the manuscript, Jonathan Pillow for helpful advice on inferential methods, and the Daw and Cohen labs for their feedback throughout. H.R. was supported by the C.V. Starr Postdoctoral Fellowship and the Connected Minds Postdoctoral Fellowship. A.J. was supported by the Google PhD Fellowship and the Wu Tsai Interdisciplinary Postdoctoral Fellowship. J.D.C. was supported by a Vannevar Bush Faculty Fellowship administered by the US Office of Naval Research. This paper is dedicated to Mark Stokes.

AUTHOR CONTRIBUTIONS

Conceptualization, H.R., N.D.D., and J.D.C.; data curation, H.R.; formal analysis, H.R. and A.J.; funding acquisition, H.R., N.D.D., and J.D.C.; methodology, H.R. and A.J.; resources, H.R., N.D.D., and J.D.C.; software, H.R.; supervision, N.D.D. and J.D.C.; writing—original draft, H.R.; writing—review & editing, H.R., A.J., N.D.D., and J.D.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Claude Code (Anthropic) for coding assistance. The authors reviewed and edited the output as needed and take full responsibility for the content of the published article.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2026.07.028.

Received: January 5, 2026
Revised: May 15, 2026
Accepted: July 13, 2026

REFERENCES

1. Monsell, S. (2003). Task switching. Trends Cogn. Sci. *7*, 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7.
2. Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A.M., and Koch, I. (2010). Control and interference in task switching—a review. Psychol. Bull. *136*, 849–874. https://doi.org/10.1037/a0019842.
3. Vandierendonck, A., Liefooghe, B., and Verbruggen, F. (2010). Task switching: interplay of reconfiguration and interference control. Psychol. Bull. *136*, 601–626. https://doi.org/10.1037/a0019791.
4. Koch, I., Poljac, E., Müller, H., and Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking-an integrative review of dual-task and task-switching research. Psychol. Bull. *144*, 557–583. https://doi.org/10.1037/bul0000144.
5. Friedman, N.P., and Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. Cortex *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023.
6. Cepeda, N.J., Kramer, A.F., and Gonzalez de Sather, J.C.M. (2001). Changes in executive control across the life span: Examination of task-switching performance. Dev. Psychol. *37*, 715–730. https://doi.org/10.1037/0012-1649.37.5.715.
7. Steyvers, M., Hawkins, G.E., Karayanidis, F., and Brown, S.D. (2019). A large-scale analysis of task switching practice effects across the lifespan. Proc. Natl. Acad. Sci. USA *116*, 17735–17740. https://doi.org/10.1073/pnas.1906788116.
8. Millan, M.J., Agid, Y., Brüne, M., Bullmore, E.T., Carter, C.S., Clayton, N.S., Connor, R., Davis, S., Deakin, B., DeRubeis, R.J., et al. (2012). Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. Nat. Rev. Drug Discov. *11*, 141–168. https://doi.org/10.1038/nrd3628.
9. Snyder, H.R. (2013). Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. Psychol. Bull. *139*, 81–132. https://doi.org/10.1037/a0028727.
10. Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task-set switching. Control of Cognitive Processes, (331–355).
11. Meiran, N. (2000). Modeling cognitive control in task-switching. Psychol. Res. *63*, 234–249. https://doi.org/10.1007/s004269900004.
12. Rogers, R.D., and Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. J. Exp. Psychol.: Gen. *124*, 207–231. https://doi.org/10.1037/0096-3445.124.2.207.

13. Monsell, S. (2017). Task set regulation. In The Wiley Handbook of Cognitive Control, *115*, T. Egner, ed. (John Wiley & Sons, Ltd), pp. 29–49. https://doi.org/10.1002/9781118920497.ch2.

14. Allport, A., Styles, E.A., and Hsieh, S.L. (1994). Shifting intentional set – exploring the dynamic control of tasks. In Attention and Performance XV: CONSCIOUS AND NONCONSCIOUS INFORMATION PROCESSING (MIT Press), pp. 421–452.

15. Wylie, G., and Allport, A. (2000). Task switching and the measurement of "switch costs". Psychol. Res. *63*, 212–233. https://doi.org/10.1007/s004269900003.

16. Werbos, P.J. (1990). Backpropagation through time: what it does and how to do it. Proc. IEEE *78*, 1550–1560. https://doi.org/10.1109/5.58337.

17. Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. J. Exp. Psychol.: Learn. Mem. Cogn. *22*, 1423–1442. https://doi.org/10.1037/0278-7393.22.6.1423.

18. Yeung, N., and Monsell, S. (2003). Switching between tasks of unequal familiarity: the role of stimulus-attribute and response-set selection. J. Exp. Psychol. Hum. Percept. Perform. *29*, 455–469. https://doi.org/10.1037/0096-1523.29.2.455.

19. Yeung, N., Nystrom, L.E., Aronson, J.A., and Cohen, J.D. (2006). Between-task competition and cognitive control in task switching. J. Neurosci. *26*, 1429–1438. https://doi.org/10.1523/JNEUROSCI.3109-05.2006.

20. Schmitz, F., and Voss, A. (2012). Decomposing task-switching costs with the diffusion model. J. Exp. Psychol. Hum. Percept. Perform. *38*, 222–250. https://doi.org/10.1037/a0026003.

21. Ruge, H., Jamadar, S., Zimmermann, U., and Karayanidis, F. (2013). The many faces of preparatory control in task switching: reviewing a decade of fMRI research. Hum. Brain Mapp. *34*, 12–35. https://doi.org/10.1002/hbm.21420.

22. Loose, L.S., Wisniewski, D., Rusconi, M., Goschke, T., and Haynes, J.-D. (2017). Switch-independent task representations in frontal and parietal cortex. J. Neurosci. *37*, 8033–8042. https://doi.org/10.1523/JNEUROSCI.3656-16.2017.

23. Qiao, L., Zhang, L., Chen, A., and Egner, T. (2017). Dynamic trial-by-trial recoding of task-set representations in the frontoparietal cortex mediates behavioral flexibility. J. Neurosci. *37*, 11037–11050. https://doi.org/10.1523/JNEUROSCI.0935-17.2017.

24. Gilbert, S.J., and Shallice, T. (2002). Task switching: a PDP model. Cogn. Psychol. *44*, 297–337. https://doi.org/10.1006/cogp.2001.0770.

25. Ardid, S., and Wang, X.-J. (2013). A tweaking principle for executive control: neuronal circuit mechanism for rule-based task switching and conflict resolution. J. Neurosci. *33*, 19504–19517. https://doi.org/10.1523/JNEUROSCI.1356-13.2013.

26. Musslick, S., Jang, S.J., Shvartsman, M., Shenhav, A., and Cohen, J.D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. Proc. Annu. Meet. Cogn. Sci. Soc. *40*.

27. Ueltzhöffer, K., Armbruster-Genç, D.J.N., and Fiebach, C.J. (2015). Stochastic dynamics underlying cognitive stability and flexibility. PLoS Comput. Biol. *11*, e1004331. https://doi.org/10.1371/journal.pcbi.1004331.

28. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179.

29. Siegel, M., Donner, T.H., and Engel, A.K. (2012). Spectral fingerprints of large-scale neuronal interactions. Nat. Rev. Neurosci. *13*, 121–134. https://doi.org/10.1038/nrn3137.

30. Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84. https://doi.org/10.1038/nature12742.

31. Cohen, J.D., Dunbar, K., and McClelland, J.L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. Psychol. Rev. *97*, 332–361. https://doi.org/10.1037/0033-295X.97.3.332.

32. Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. (2001). Conflict monitoring and cognitive control. Psychol. Rev. *108*, 624–652. https://doi.org/10.1037/0033-295X.108.3.624.

33. Sussillo, D., and Barak, O. (2013). Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Comput. *25*, 626–649. https://doi.org/10.1162/NECO_a_00409.

34. Soldado-Magraner, J., Mante, V., and Sahani, M. (2024). Inferring context-dependent computations through linear approximations of prefrontal cortex dynamics. Sci. Adv. *10*, eadl4743. https://doi.org/10.1126/sciadv.adl4743.

35. Gurnani, H., Liu, W., and Brunton, B.W. (2024). Feedback control of recurrent dynamics constrains learning timescales during motor adaptation. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.24.595772.

36. Brown, E.N., Frank, L.M., Tang, D., Quirk, M.C., and Wilson, M.A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. J. Neurosci. *18*, 7411–7425. https://doi.org/10.1523/JNEUROSCI.18-18-07411.1998.

37. Smith, A.C., and Brown, E.N. (2003). Estimating a state-space model from point process observations. Neural Comput. *15*, 965–991. https://doi.org/10.1162/089976603765202622.

38. Macke, J.H., Büsing, L., Cunningham, J.P., Yu, B.M., Shenoy, K.V., and Sahani, M. (2011). Empirical models of spiking in neural populations. Adv. Neural Inf. Process. Syst. *24*, 1350–1358.

39. LaFosse, P.K., Zhou, Z., O'Rawe, J.F., Friedman, N.G., Scott, V.M., Deng, Y., and Histed, M.H. (2024). Single-Cell Optogenetics Reveals Attenuation-by-Suppression in Visual Cortical Neurons. Proc. Natl. Acad. Sci *121*, e2318837121. https://doi.org/10.1073/pnas.2318837121.

40. Nozari, E., Bertolero, M.A., Stiso, J., Caciagli, L., Cornblath, E.J., He, X., Mahadevan, A.S., Pappas, G.J., and Bassett, D.S. (2023). Macroscopic resting-state brain dynamics are best described by linear models. Nat. Biomed. Eng. *8*, 68–84. https://doi.org/10.1038/s41551-023-01117-y.

41. Jha, A., Gupta, D., Brody, C.D., and Pillow, J.W. (2024). Disentangling the roles of distinct cell classes with cell-type dynamical systems. Adv. Neural. Inf. Process Syst. *37*, 33668–33690. https://doi.org/10.52202/079017-1060.

42. Roweis, S., and Ghahramani, Z. (1999). A unifying review of linear gaussian models. Neural Comput. *11*, 305–345. https://doi.org/10.1162/089976699300016674.

43. Brunton, S.L., Budišić, M., Kaiser, E., and Kutz, J.N. (2022). Modern Koopman theory for dynamical systems. SIAM Rev. *64*, 229–340. https://doi.org/10.1137/21M1401243.

44. Möcks, J. (1986). The influence of latency jitter in principal component analysis of event-related potentials. Psychophysiology *23*, 480–484. https://doi.org/10.1111/j.1469-8986.1986.tb00659.x.

45. Williams, R.L., and Lawrence, D.A. (2007). Linear State-Space Control Systems (John Wiley & Sons). https://doi.org/10.1002/9780470117873.

46. Tang, E., and Bassett, D.S. (2018). Colloquium: Control of dynamics in brain networks. Rev. Mod. Phys. *90*. https://doi.org/10.1103/RevModPhys.90.031003.

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

48. Gohil, C., Kohl, O., Huang, R., van Es, M.W.J., Parker Jones, O., Hunt, L.T., Quinn, A.J., and Woolrich, M.W. (2024). Dynamic network analysis of electrophysiological task data. Imaging Neurosci. (Camb) *2*. imag-2-00226. https://doi.org/10.1162/imag_a_00226.

49. Misra, J., and Pessoa, L. (2026). Human brain dynamics and spatiotemporal trajectories during threat processing. eLife *14*, RP102539. https://doi.org/10.7554/eLife.102539.

50. Hess, F., Monfared, Z., Brenner, M., and Durstewitz, D. (2023). Generalized teacher forcing for learning chaotic dynamics. In International Conference on Machine Learning (PMLR), pp. 13017–13049.

51. Pals, M., Sagtekin, A.E., Pei, F., Gloeckler, M., and Macke, J.H. (2024). Inferring stochastic low-rank recurrent neural networks from neural data. In Advances in Neural Information Processing Systems 37 (Neural Information Processing Systems Foundation, Inc. (NeurIPS)), pp. 18225–18264. https://doi.org/10.52202/079017-0579.

52. Ghahramani, Z. and Hinton, G.E. (1996) Parameter Estimation for Linear Dynamical Systems. Technical Report CRG-TR-96-2.

53. Murphy, K.P. (2023). Probabilistic Machine Learning: Advanced Topics (MIT Press).

54. Ritz, H. (2024). harrisonritz/StateSpaceAnalysis.jl v0.2.0. Zenodo. https://doi.org/10.5281/zenodo.14511205.

55. Brunton, S.L., Brunton, B.W., Proctor, J.L., Kaiser, E., and Kutz, J.N. (2017). Chaos as an intermittently forced linear system. Nat. Commun. *8*, 19. https://doi.org/10.1038/s41467-017-00030-8.

56. Hall-McMaster, S., Muhle-Karbe, P.S., Myers, N.E., and Stokes, M.G. (2019). Reward boosts neural coding of task rules to optimize cognitive flexibility. J. Neurosci. *39*, 8549–8561. https://doi.org/10.1523/JNEUROSCI.0631-19.2019.

57. Arnau, S., Liegel, N., and Wascher, E. (2024). Frontal midline theta power during the cue-target-interval reflects increased cognitive effort in rewarded task-switching. Cortex *180*, 94–110. https://doi.org/10.1016/j.cortex.2024.08.004.

58. Kao, T.-C., and Hennequin, G. (2019). Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics. Curr. Opin. Neurobiol. *58*, 122–129. https://doi.org/10.1016/j.conb.2019.09.001.

59. Holroyd, C.B. (2024). The controllosphere: The neural origin of cognitive effort. Psychol. Rev. *132*, 603–631. https://doi.org/10.1037/rev0000467.

60. Klett, C., Abate, M., Yoon, Y., Coogan, S., and Feron, E. (2020). Bounding the state covariance matrix for switched linear systems with noise. In American Control Conference (ACC) (Institute of Electrical and Electronics Engineers), pp. 2876–2881. https://doi.org/10.23919/ACC45564.2020.9147787.

61. King, J.A., Korb, F.M., von Cramon, D.Y., and Ullsperger, M. (2010). Post-error behavioral adjustments are facilitated by activation and suppression of task-relevant and task-irrelevant information processing. J. Neurosci. *30*, 12759–12769. https://doi.org/10.1523/JNEUROSCI.3274-10.2010.

62. Luyckx, F., Nili, H., Spitzer, B., and Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning. eLife *8*, e42816. https://doi.org/10.7554/eLife.42816.

63. Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. Preprint at arXiv. arXiv:1412.3555. https://doi.org/10.48550/arXiv.1412.3555

64. Can, T., Krishnamurthy, K., and Schwab, D.J. (2020). Gating creates slow modes and controls phase-space complexity in GRUs and LSTMs. Proc. Mach. Learn. Res. *107*, 476–511. https://proceedings.mlr.press/v107/can20a.html.

65. Braver, T.S., and Cohen, J.D. (1999). Dopamine, cognitive control, and schizophrenia: the gating model. Prog. Brain Res. *121*, 327–349. https://doi.org/10.1016/S0079-6123(08)63082-4.

66. O'Reilly, R.C., and Frank, M.J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput. *18*, 283–328. https://doi.org/10.1162/089976606775093909.

67. Marder, E., and Taylor, A.L. (2011). Multiple models to capture the variability in biological neurons and networks. Nat. Neurosci. *14*, 133–138. https://doi.org/10.1038/nn.2735.

68. Pagan, M., Tang, V.D., Aoi, M.C., Pillow, J.W., Mante, V., Sussillo, D., and Brody, C.D. (2024). Individual variability of neural computations underlying flexible decisions. Nature *639*, 421–429. https://doi.org/10.1038/s41586-024-08433-6.

69. Huang, A., Singh, S.H., Martinelli, F., and Rajan, K. (2025). Measuring and Controlling Solution Degeneracy across Task-Trained Recurrent Neural Networks. In The Thirty-Ninth Annual Conference on Neural Information Processing Systems.. https://openreview.net/forum?id=R0LqbSgZjP.

70. Song, H., Shim, W.M., and Rosenberg, M.D. (2023). Large-scale neural dynamics in a shared low-dimensional state space reflect cognitive and attentional dynamics. Elife *12*, e85487. https://doi.org/10.7554/eLife.85487.

71. Athalye, V.R., Carmena, J.M., and Costa, R.M. (2019). Neural reinforcement: re-entering and refining neural dynamics leading to desirable outcomes. Curr. Opin. Neurobiol. *60*, 145–154. https://doi.org/10.1016/j.conb.2019.11.023.

72. Ritz, H., Leng, X., and Shenhav, A. (2022). Cognitive control as a multivariate optimization problem. J. Cogn. Neurosci. *34*, 569–591. https://doi.org/10.1162/jocn_a_01822.

73. Badre, D., and Wagner, A.D. (2006). Computational and neurobiological mechanisms underlying cognitive flexibility. Proc. Natl. Acad. Sci. USA *103*, 7186–7191. https://doi.org/10.1073/pnas.0509550103.

74. Jaffe, P.I., Poldrack, R.A., Schafer, R.J., and Bissett, P.G. (2023). Modelling human behaviour in cognitive tasks with latent dynamical systems. Nat. Hum. Behav. *7*, 986–1000. https://doi.org/10.1038/s41562-022-01510-8.

75. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv. Neural. Inf. Process. Syst. *32*, 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

76. Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. *2011*, 156869. https://doi.org/10.1155/2011/156869.

77. Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods *134*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

78. Crameri, F., Shephard, G.E., and Heron, P.J. (2020). The misuse of colour in science communication. Nat. Commun. *11*, 5444. https://doi.org/10.1038/s41467-020-19160-7.

79. Loshchilov, I., and Hutter, F. (2017) Decoupled Weight Decay Regularization.

80. Ehrlich, D.B., and Murray, J.D. (2022). Geometry of neural computation unifies working memory and planning. Proc. Natl. Acad. Sci. USA *119*, e2115610119. https://doi.org/10.1073/pnas.2115610119.

81. Langdon, C., and Engel, T.A. (2025). Latent circuit inference from heterogeneous neural responses during cognitive tasks. Nat. Neurosci. *28*, 665–675. https://doi.org/10.1038/s41593-025-01869-7.

82. Linderman, S.W., Chang, P., Harper-Donnelly, G., Kara, A., Li, X., Duran-Martin, G., and Murphy, K. (2025). Dynamax: A python package for probabilistic state space modeling with JAX. J. Open Source Softw. *10*, 7069. https://doi.org/10.21105/joss.07069.

83. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). Springer.

84. Holmes, E., Ward, E., and Wills, K. (2012). MARSS: Multivariate autoregressive state-space models for analyzing time-series data. R J. *4*, 11. https://doi.org/10.32614/RJ-2012-002.

85. Särkkä, S. (2013). *Bayesian Filtering and Smoothing* (Cambridge University Press). https://doi.org/10.1017/CBO9781139344203.

86. Smith, G., Freitas, J.F.G., Robinson, T., and Niranjan, M. (1999). Speech modelling using subspace and EM techniques. Neural Inf. Process. Sys. 796–802. https://papers.nips.cc/paper_files/paper/1999/file/d860bd12ce9c026814bbdfc1c573f0f5-Paper.pdf.

87. Stone, I.R., Sagiv, Y., Park, I.M., and Pillow, J.W. (2023). Spectral learning of Bernoulli linear dynamical systems models for decision-making. Trans. Mach. Learn. Res. *982*. https://openreview.net/forum?id=giw2vcAhiH.

88. Larimore, W.E. (1990). Canonical variate analysis in identification, filtering, and adaptive control. In 29th IEEE Conference on Decision and Control. https://doi.org/10.1109/CDC.1990.203665.

89. Huang, A., Ostrow, M., Singh, S.H., Kozachkov, L., Fiete, I.R., and Rajan, K. (2026). InputDSA: Demixing, then comparing recurrent and externally driven dynamics. The Fourteenth International Conference on Learning Representations. https://openreview.net/forum?id=2FrAvRz1E7.

90. Tallon-Baudry, C., Bertrand, O., Delpuech, C., and Pernier, J. (1996). Stimulus specificity of phase-locked and non-phase-locked 40 hz visual responses in human. J. Neurosci. *16*, 4240–4249. https://doi.org/10.1523/JNEUROSCI.16-13-04240.1996.

91. Valente, A., Ostojic, S., and Pillow, J.W. (2022). Probing the relationship between latent linear dynamical systems and low-rank recurrent neural network models. Neural Comput. *34*, 1871–1892. https://doi.org/10.1162/neco_a_01522.

92. Churchland, M.M., and Shenoy, K.V. (2024). Preparatory activity and the expansive null-space. Nat. Rev. Neurosci. *25*, 213–236. https://doi.org/10.1038/s41583-024-00796-z.

93. Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. Neuroimage *46*, 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025.

94. Ehinger, B.V., and Dimigen, O. (2019). Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. PeerJ *7*, e7838. https://doi.org/10.7717/peerj.7838.

95. Cox, D.R., and Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. Am. Stat. *46*, 1–4. https://doi.org/10.1080/00031305.1992.10475836.

96. Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507. https://doi.org/10.1126/science.1127647.

97. Staffans, O. (2009). Encyclopedia of Mathematics and Its Applications Well-Posed Linear Systems Series Number 103 (Cambridge University Press).

98. Kalman, R.E. (2010). Lectures on controllability and observability. In Controllability and Observability, E. Evangelisti, ed. (Springer), pp. 1–149. https://doi.org/10.1007/978-3-642-11063-4_1.

99. King, J.R., and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn. Sci. *18*, 203–210. https://doi.org/10.1016/j.tics.2014.01.002.

100. Smith, S.M., and Nichols, T.E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage *44*, 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061.

101. Mensen, A., and Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. Neuroimage *67*, 111–118. https://doi.org/10.1016/j.neuroimage.2012.10.027.

102. Zaykin, D.V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J. Evol. Biol. *24*, 1836–1841. https://doi.org/10.1111/j.1420-9101.2011.02297.x.

103. Lohmiller, W., and Slotine, J.-J.E. (1998). On contraction analysis for nonlinear systems. Automatica *34*, 683–696. https://doi.org/10.1016/S0005-1098(98)00019-3.

104. Jaffe, S., Davydov, A., Lapsekili, D., Singh, A., and Bullo, F. (2024). Learning neural contracting dynamics: Extended linearization and global guarantees. In Advances in Neural Information Processing Systems 37 (Neural Information Processing Systems Foundation, Inc. (NeurIPS)), pp. 66204–66225. https://doi.org/10.52202/079017-2116.

105. Elman, J.L. (1990). Finding structure in time. Cogn. Sci. *14*, 179–211. https://doi.org/10.1207/s15516709cog1402_1.

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
'Hall-McMaster' EEG dataset	Hall-McMaster et al.[56]	www.osf.io/kuzye/
'Arnau' EEG dataset	Arnau, Liegel, & Wascher[57]	www.osf.io/ndgst/
Software and algorithms		
Study-specific analysis code	This Paper	www.github.com/harrisonritz/convergent-neural-dynamical-systems
State-space analysis package	This Paper	https://doi.org/10.5281/zenodo.14511205
PyTorch (v2.4.0)	Paszke et al.[75]	www.pytorch.org
MatlabTFCE	Mark Thornton	www.github.com/markallenthornton/MatlabTFCE
ControlSystemIdentification.jl (v2.10.2)	Fredrik Bagge Carlson	www.github.com/baggepinnen/ControlSystemIdentification.jl
FieldTrip	Oostenveld, Fries, Maris, & Schoffelen[76]	https://www.fieldtriptoolbox.org
EEGLAB (v2024.2)	Delorme & Makeig[77]	https://sccn.ucsd.edu/eeglab/
Scientific Color Maps	Crameri, Shephard, & Heron[78]	https://www.fabiocrameri.ch/colourmaps/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

EEG datasets

Hall-McMaster sample

The first EEG dataset was originally published by Hall-McMaster and colleagues,[56] and was made open-access at www.osf.io/kuzye/. Thirty participants (18–35 years old; 19 females) with normal or corrected-to-normal vision and no history of neurological or psychological disorders underwent an EEG recording session during task performance. This study was approved by the Central University Research Ethics Committee at the University of Oxford, with all participants providing informed consent. See [56] for the full description.

Arnau sample

The second EEG dataset was originally published by Arnau and colleagues,[57] and was made open-access at www.osf.io/ndgst/. This dataset consists of twenty-six participants (23 female, mean(SD) age 21.65(2.15) years), with normal or corrected-to-normal vision and no history of neurological or psychological disorders, and verified normal color vision. This study was approved by the ethics committee of the Leibniz Research Centre for Working Environment and Human Factors, Dortmund, with all participants providing informed consent. See [57] for the full description.

METHOD DETAILS

Task-optimized recurrent neural networks (RNNs)

RNN architecture

We trained RNNs with gated recurrent units (GRUs; 108 hidden units; [28]) to perform task-switching. These networks used a standard single-layer RNN architecture: linearly encoded inputs (u_t), a recurrent hidden state (x_t) with a hyperbolic tangent activation function, and linearly decoded outputs (y_t) with a sigmoid link function. Note that for consistency with the SSM, here we use x_t to refer to the latent state (PyTorch: h_t), and u_t to refer to the inputs (PyTorch: x_t).

We trained the RNNs using backpropagation through time (learning rule: AdamW,[79] learning rate =.01, weight decay =.01). The GRU component of this RNN learned two 'gates', each of which encodes the hidden state and inputs, and outputs a sigmoid-constrained gating value that is element-wise multiplied with the hidden state. The 'reset gate' (r_t) is applied to the hidden state before the activation function, producing the 'new state' (n_t). The 'update gate' (z_t) is applied after the activation function, mixing the new state and the previous hidden state. This was implemented through the standard GRU class in PyTorch[75]:

$$r_t = \sigma(W_{hr}x_{t-1} + b_{hr} + W_{ir}u_t + b_{ir})$$

$$n_t = \tanh(r_t \odot (W_{hn}x_{t-1} + b_{hn}) + W_{in}u_t + b_{in})$$

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

$$x_t = (1 - z_t) \odot n_t + z_t \odot x_{t-1}$$

$$y_t = \sigma(W_{ho} x_t)$$

RNN curricula

We trained four groups of GRUs on different curricula (512 RNNs per curriculum, matched random seeds), crossing two levels of switch-training' against two levels of trial-spacing'.

Switch-training

'1-Trial' RNNs were trained on sequences containing a single trial (9600 sequences × 500 epochs; see task description below). '2-Trial' RNNs were trained on sequences of one trial for 90% of their training (450 epochs), and then trained on sequences of two trials for the final 10% of training (50 epochs). Each epoch had an equal number of target transitions, distractor transitions, and task transitions (64 conditions × 150 repetitions). The rationale for this design was to teach both groups how to perform the task, but give 2-Trial networks enough experience that they would learn compensatory dynamics for switching. By 450 epochs, RNNs were over-trained on the task, with increasing test loss (Figure 1E).

Trial-spacing

'Short ITI' RNNs were trained on sequences in which the inter-trial interval between trials was short (20 timesteps; see task description below), whereas 'Long ITI' RNNs were trained on sequences with longer ITIs (60 timesteps). This ratio of ITIs was chosen to match the EEG experiments (see below). 2-Trial RNNs were tested on the same ITI duration on which they were trained. For 1-Trial RNNs, those in the Short-ITI condition were tested using short ITIs, and those in the Long-ITI condition were tested using long ITIs.

RNN task

RNNs performed an extension of the 'context-dependent decision making' task[29,30,80,81] On one-trial sequences, RNNs performed the standard version of this task. In each sequence, RNNs first received a task cue (one-hot coded; 10 timesteps), followed by a delay period (20 timesteps). During the task period, they received two pairs of stimulus inputs (40 timesteps). They had to judge which input in the task-relevant pair (i.e., as indicated by the task cue) had a higher amplitude. On two-trial sequences, RNNs performed two consecutive trials (i.e., without resetting the state or gradient). They performed the first trial, had an ITI with no inputs (20 or 60 timesteps, see above), and then performed the second trial.

RNNs were trained to optimize a binary cross-entropy loss function on the correct ($y_i = 1$) and incorrect ($y_i = 0$) response options ($\mathcal{L} = -\sum_i (y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i))$), with the loss starting 10 timesteps after the onset of the stimuli to encourage evidence integration. There were no other training targets during the epoch. We added normally distributed noise to the inputs in each sequence (cue SNR: 0 dB; trial SNR: -16.5 dB), which was resampled every epoch. Noise was added to every timestep of the sequence, encouraging robust task dynamics even during inter-trial and inter-stimulus intervals. We trained at relatively low SNRs to encourage the networks to use task control and discourage overfitting. The test loss was evaluated on the 64 conditions, presented without noise. To provide a stronger test of generalization, we also evaluated the trained RNNs on novel three-trial sequences (Figures 1F and 1G).

Hidden unit visualization

We used singular value decomposition (SVD) to visualize hidden unit activation during two-trial sequences. SVD is a method for performing PCA. For PCA, however, the columns of the data matrix are first mean-centered, producing a decomposition of the covariance. We chose SVD rather than PCA because RNNs had a common and interpretable zero point (RNNs were initialized at the origin), so mean-centering would needlessly complicate the comparisons between the networks.

Using trained networks, we first simulated 2048 noisy trials for an example network from each curriculum (same random seed). Next, we concatenated the hidden unit activations across sequences and networks into a 2D matrix $((\text{timesteps} \times \text{sequences}) \times (\text{hidden units} \times \text{curricula}))$. Performing SVD on this matrix provided embeddings that captured the shared and distinct components of each network, which were qualitatively similar to when we used a single embedding for all RNNs. We then averaged the sequences within each network and task transitions, and projected these average sequences onto the right singular vectors corresponding to each network.

EEG datasets

Hall-McMaster dataset

Task. Participants performed a cued task-switching experiment during the EEG recording. A central goal of the original study was to investigate the role of incentives on task encoding, a manipulation that was not the focus of the current experiment.

Participants saw a cue for high vs. low rewards for correct performance (800 ms, 400 ms delay), a cue for the shape vs. color task (200 ms, 400 ms delay), a task stimulus consisting of a square or circle colored in yellow or blue (min(RT, 1400 ms)), reward feedback (200 ms), and an inter-trial interval (1000–1400 ms; response-cue interval: 2400–2800 ms). Participants' task was to respond to the trial stimulus according to the cued rule: if the task was 'shape', respond with a key press based on the shape; if the task was 'color', respond with a key press based on the color. Participants practiced the task until they reached a 70% accuracy criterion, and then performed 10 blocks of 65 trials in the main task. Conditions were balanced within each block.

EEG preprocessing. For our analyses, we included trials 1) that were not the first trial of a block, 2) where the current and previous trial were accurate, 3) where the current RT was longer than 200 ms, and 4) that were not identified in the original experiment as containing artifacts during preprocessing. We epoched the data between the onset of the task cue and 200 ms into the task. To avoid removing previous task representations, we did not apply any baseline correction.

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

Working with the preprocessed data from the original experiment, we low-pass filtered the EEG at 30 Hz (resulting in a 0.01 Hz–30 Hz bandpass filter) and then resampled the data to 125 Hz using FIELDTRIP. We split the experimental blocks into training and test sets, with Mean(SD) = 416(63) training trials and 47(9) testing trials. EEG data were recorded with 61 Ag/AgCl sintered electrodes (EasyCap; 10-10 layout (FIELDTRIP template: elec1010)), a NeuroScan SynAmps RT amplifier, and Curry 7 acquisition software.

Arnau dataset

Task

Participants performed a cued task-switching experiment during the EEG recording. On each trial, participants saw a cue for the tilt vs. color task (200 ms, 600 ms delay). During the task phase, participants saw an array of Gabor patches and had to report either the tilt or the color of the singleton, depending on the cued task. Participants had up to 1200 ms to respond, and then there was an inter-trial interval (800–1000 ms). As in Hall-McMaster, participants practiced the task before the main experiment, which consisted of eight experimental blocks with 256 trials each. This experiment also investigated the role of incentives on task switching using a mini-block-level incentive manipulation. This similarity is coincidental, due to motivation being an active area of focus in recent experiments on cognitive control in general and task switching in particular.

EEG preprocessing

EEG data were recorded using 128 passive Ag/AgCl electrodes (Easycap GmbH, Herrsching, Germany) with a NeurOne Tesla AC-amplifier (Bittium Biosignals Ltd., Kuopio, Finland). We used the same performance-based inclusion criteria as in Hall-McMaster. EEG data were preprocessed using EEGLab scripts modified from the original experiment. We high-pass filtered at 0.1 Hz and low-pass filtered at 30 Hz using FIR filters, and then resampled the data at 125 Hz. To avoid removing previous task representations, we did not apply any baseline correction. We used more lenient inclusion criteria for the ICA auto-rejection procedure than the original experiment, as we found that some participants had most of their components removed. After preprocessing, we split up the experimental blocks into training and test sets, with Mean(SD) = 676(147) training trials and 142(31) testing trials.

QUANTIFICATION AND STATISTICAL ANALYSIS

Linear-Gaussian state-space model

Generative model

We developed a toolkit called STATESPACEANALYSIS for fitting linear-Gaussian state-space models (SSMs) to neuroimaging data, inspired by the SSM and DYNAMAX packages in PYTHON.[82] Our goal was to test how effectively SSMs could capture rich neuroimaging data, balancing the predictive power of machine learning approaches with the interpretability offered by linear-Gaussian system identification. To provide the speed and accuracy needed for fitting high-dimensional latent variable models, we developed this package using the powerful open-source coding language JULIA, available at www.github.com/harrisonritz/StateSpaceAnalysis.jl.[54]

The generative model for the analysis is a partially observable autoregressive process: a linear dynamical system from which we can only make noisy measurements (or, equivalently, that only produces noisy emissions of its underlying state). This autoregressive process consists of a vector of latent factors ($\boldsymbol{x}$), and a discrete-time difference equation that describes how they evolve:

$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + B\boldsymbol{u}_{t-1} + \boldsymbol{w}_t$$

$$\boldsymbol{x}_{t=1} = B_0\boldsymbol{u}_0 + \boldsymbol{w}_1$$

Latent factors evolve according to their recurrent dynamics ($A\boldsymbol{x}$), the influence of known inputs ($B\boldsymbol{u}$), and process noise ($\boldsymbol{w}_t \sim \mathcal{N}(0, W)$). The initial conditions depend on trial conditions ($B_0\boldsymbol{u}_0$) and initial uncertainty ($\boldsymbol{w}_1 \sim \mathcal{N}(0, W_1)$).

At each timestep, we get a noisy observation of these latent factors:

$$\boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{v}_t$$

Observations are projections of the latent factors ($C\boldsymbol{x}$) corrupted by observation noise ($\boldsymbol{v}_t \sim \mathcal{N}(0, V)$). Note that only process noise is carried forward in time through the recurrent dynamics, unlike observation noise.

To provide more flexibility to our input model, we modeled time-varying inputs with a set of cubic B-splines ('basis splines'; Figure S2). B-splines are piecewise polynomial functions that provide local approximations to smooth functions.[83] This smoothness constraint can help capture realistically time-varying signals while reducing overfitting. Particularly for EEG, these flexible inputs will help capture neural computations poorly captured by our recurrent dynamics (e.g., below the spatial resolution of EEG), while staying within our linear framework. Spline knots were placed to approximately optimize coverage over the epoch (using the averagebasis procedure in BSplines.jl).

Expectation maximization

One benefit of working with linear-Gaussian SSMs is that the marginal data likelihood ($P(\boldsymbol{y}_{1:T}|\Theta)$) can be computed analytically. While this marginal likelihood would allow us to directly fit the parameters (Θ) using techniques like maximum likelihood estimation, in practice it is often more convenient to use the expectation-maximization algorithm (EM; [52,53,84]). EM maximizes a lower bound on the marginal log likelihood; adding priors makes this a lower bound on the log posterior. EM alternates between an Expectation Step (E-step) and a Maximization Step (M-step), each of which has an analytic solution for linear-Gaussian models. Because the E-step makes this bound tight at the current parameter estimates, EM is guaranteed not to decrease this objective at each iteration.

The E-step computes the posterior over the latent states using the Rauch–Tung–Striebel (RTS) smoother: $(P(x_{\{t\}}|y_{\{1:T\}}))$. This smoother extends the Kalman filter, which conditions only on the current and previous observations: $(P(x_{\{t\}}|y_{\{1:t\}}))$. The M-step uses the smoothed latents to find the model parameters that maximize the expected complete-data log-posterior (see below). The linear-Gaussian assumptions considerably simplify this procedure: the M-step depends on the smoothed latents only through their first and second moments (their means, covariances, and lag-one cross-covariances), and can be solved analytically using Bayesian multivariate linear regression.

Expected complete-data log-posterior

The M-step maximizes the expected complete-data log-posterior over N epochs and T timesteps:

$$\mathcal{L} = \mathbb{E}\left[\sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_1^n; B_0\boldsymbol{u}_0, W_1)\right] + \log \mathcal{MN}\left(B_0; 0, W_1, \Lambda_{B_0}^{-1}\right) +$$

$$\mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=2}^{T} \log \mathcal{N}(\boldsymbol{x}_t^n; A\boldsymbol{x}_{t-1}^n + B\boldsymbol{u}_{t-1}^n, W)\right] + \log \mathcal{MN}(A; 0, W, \Lambda_A^{-1}) + \log \mathcal{MN}(B; 0, W, \Lambda_B^{-1}) +$$

$$\mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=1}^{T} \log \mathcal{N}(\boldsymbol{y}_t^n; C\boldsymbol{x}_t^n, V)\right] + \log \mathcal{MN}(C; 0, V, \Lambda_C^{-1})$$

We used weakly informative priors for $[A, B, B_0, C]$, with prior means of zero and prior column precisions set to scaled identity matrices ($\Lambda = 10^{-6}I$). We used uninformative priors for $[W, W_1, V]$, whose log-prior terms we omit above.

E-step

Using the filter-smoother algorithm,[85] we estimated the posterior mean ($\boldsymbol{m}_t$) and covariance (Σ_t) of the latent state across N epochs and T timesteps, generating a set of sufficient statistics:

$$M(t_1, t_2) = \sum_{n=1}^{N}\sum_{t=t_1}^{t2} m_{n,t}m_{n,t}^\top + \sum\nolimits_t \qquad U_u = \sum_{n=1}^{N}\sum_{t=1}^{T-1} u_{n,t}u_{n,t}^\top$$

$$M_\Delta = \sum_{n=1}^{N}\sum_{t=1}^{T-1} m_{n,t+1}m_{n,t}^\top + \sum\nolimits_{t+1,t} \qquad U_m = \sum_{n=1}^{N}\sum_{t=1}^{T-1} u_{n,t}m_{n,t}^\top$$

$$U_\Delta = \sum_{n=1}^{N}\sum_{t=1}^{T-1} m_{n,t+1}u_{n,t}^\top$$

$$Y_y = \sum_{n=1}^{N}\sum_{t=1}^{T} y_{n,t}y_{n,t}^\top \qquad U_{u_0} = \sum_{n=1}^{N} u_{n,0}u_{n,0}^\top$$

$$Y_\Delta = \sum_{n=1}^{N}\sum_{t=1}^{T} y_{n,t}m_{n,t}^\top \qquad U_{\Delta_0} = \sum_{n=1}^{N} m_{n,1}u_{n,0}^\top$$

M-step

We then use the computed sufficient statistics to update the parameters using a maximization procedure similar to ordinary least squares, with the inclusion of ridge priors on all of the dynamics matrices (Λ).

Dynamics terms:

$$[A \quad B] \leftarrow [M_\Delta \quad U_\Delta] \begin{bmatrix} M_{(1,T-1)}+\Lambda_A & U_m^\top \\ U_m & U_u+\Lambda_B \end{bmatrix}^{-1}$$

$$C \leftarrow Y_\Delta \left(M_{(1,T)}+\Lambda_C\right)^{-1}$$

$$B_0 \leftarrow U_{\Delta_0}\left(U_{u_0}+\Lambda_{B_0}\right)^{-1}$$

Covariance terms:

$$W \leftarrow \frac{1}{N(T-1) - d_x}\Big(M_{(2,T)} + AM_{(1,T-1)}A^\top + BU_uB^\top$$

$$- M_\Delta A^\top - A\,M_\Delta^\top - U_\Delta B^\top - B\,U_\Delta^\top$$

$$+ AU_m^\top B^\top + BU_mA^\top + A\Lambda_A A^\top + B\Lambda_B B^\top\Big)$$

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

$$V \leftarrow \frac{1}{NT - d_y}\left(Y_y + CM_{(1,T)}C^\top - Y_\Delta C^\top - CY_\Delta^\top + C\Lambda_C C^\top\right)$$

$$W_1 \leftarrow \frac{1}{N - d_x}\left(M_{(1,1)} + B_0 U_{u_0} B_0^\top - U_{\Delta_0} B_0^\top - B_0 U_{\Delta_0}^\top + B_0 \Lambda_{B_0} B_0^\top\right)$$

With d_x and d_y indicating the number of factors and the number of observation dimensions, respectively. Note that the estimator of W_1 is unusual, as the initial conditions are not usually parameterized with inputs; however, this is equivalent to the standard formula[52] when the only input is an intercept. Alternating between these E-steps and M-steps monotonically improves the expected complete-data log-posterior towards a local optimum, with a strong dependence on the initialization (see 'Subspace Identification' below).

Subspace identification (SSID)

We found that good initialization of the parameters was critical to the effectiveness and stability of the model-fitting procedure. We initialized the parameters using a procedure called subspace identification (SSID; [86,87]). This procedure constructs a large delay embedding matrix which concatenates lagged future and past copies of the observations (***y***) and inputs (***u***). The 'horizon' of these lags was set to match the number of factors in the largest tested model (see below). We used the canonical variate analysis procedure,[88] which uses canonical correlation analysis to find low-dimensional mappings between lagged inputs and outputs. We can recover the *A* and *C* matrices by processing different blocks of the canonical variate matrices, and then estimate the rest of the parameters using infinite-horizon Kalman filtering. We performed canonical variate analysis using a modified version of the CONTROLSYSTEMIDENTIFICATION.JL Julia package (STAR Methods).

Benefits of state-space modeling

State-space models offer several advantages over sensor level analyses (e.g., regression) or spatial decompositions (e.g., PCA).

First, SSMs can approximately linearize a nonlinear system by lifting the dimensionality. This rationale comes from Koopman operator theory, in which a nonlinear dynamical system can be globally linearized by embedding the system in a (infinitely) high-dimensional space.[43] A popular approximation to the Koopman operator is 'Dynamic mode decomposition' (DMD), which uses decompositions of delay-embedding matrices to create linearizing embeddings.[55] Our estimation of time-varying inputs using our spline bases may help provide the 'intermittent inputs' that have been found to help linearization in DMD.[55] SSMs offer a similar embedding to DMD, albeit with more structure (e.g., the lagged influence of previous measurements depends on the latent recurrent dynamics). Indeed, our parameter initialization, based on Subspace Identification, is also a decomposition of delay embedding matrices. From a recent review of Koopman operator theory: 'It is important to note that these subspace system identification approaches are designed to model linear systems, and it is unclear how to interpret the results when they are applied to nonlinear systems. However, the strong connection between DMD and the Koopman operator provides a valuable new perspective on how to interpret these approaches, even when used to model data from nonlinear dynamical systems.'[43] In practice, we found that SSMs had a high degree of predictive accuracy, outperforming nonlinear RNNs for predicting EEG (Figure 3), suggesting that this approximate linearization was a reasonable characterization of the system.

Approximate linearization has several benefits. First, without any of the linearization provided by SSMs (e.g., only performing PCA), nonlinear manifolds in RNNs or EEG can distort similarity metrics based on angles or Euclidean distances (e.g., [89]). Second, the linear structure of SSMs allows us to isolate the contributions of different inputs (e.g., task state analyses; STAR Methods). Without this linearization, the effects of inputs will depend on the state, complicating the interpretation of input-driven dynamics.

Second, unlike PCA or encoding analyses, SSMs explicitly model the system's dynamics. Whereas standard encoding methods capture the instantaneous effects of task conditions, SSMs can capture how this instantaneous encoding is propagated over time. This can provide an estimate of unique inputs at each timestep (i.e., controlling for the recurrent carry-over of previous inputs), and can dissociate recurrent and input-driven contributions to latent states. While these features were not critical for the current experiment, they may benefit future experiments, for example by linking recurrent dynamics to inter-regional connectivity.

Third, SSMs are probabilistic models, allowing them to better accommodate trial-to-trial temporal jitter and drift than methods like PCA or regression.[44] SSMs can capture these sources of noise through the initial state uncertainty (W_1), the process noise (W), and how noise propagates through the recurrent dynamics (A). By estimating the SSM parameters based on posterior estimates of the latent state, this fitting procedure should be more robust to these between-trial variations. While SSM inputs are time-locked, the spline bases can estimate smooth 'evoked' responses, whereas noise and recurrent dynamics may capture additional 'induced' responses.[90]

Finally, SSMs were highly predictive models of neural activity. We compared SSMs to an autoregressive encoding model fit directly to the observation components (with vector ('all-to-all') autoregression and matched inputs). SSMs had 10× better predictive accuracy in held-out EEG data (i.e., $R^2 = .90$ relative to the autoregressive encoding model; Figure 3B), demonstrating that SSMs are capturing more than just the temporal smoothness of the data. Unlike AR(1) models, SSMs can capture non-Markovian (i.e., 'long-range') relationships through their latent dynamics.[91,92] Thus, SSMs offer a more accurate surrogate model than PCA and/or regression, allowing for higher confidence when interpreting the inferred latent dynamics as reflections of the neural recordings (especially in conjunction with the accurate parameter recovery; Figures 2D and S3).

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

Fitting procedure

SSM structure and preprocessing: RNN

The fitting procedure was closely matched between networks and EEG. We separately preprocessed RNN training and test trials (training: 1536 trials; testing: 512) by performing PCA on the hidden unit trajectories (keeping components accounting for 99% of the variance; estimated on training data only) and constructing a temporal basis for the inputs (10 spline bases over the epoch). Inputs included a (time-varying) bias, the cued task, the previous trial's task, switch vs. repeat. Inputs to initial conditions were an intercept and the previous task. Each input was z-scored across trials to standardize and reduce collinearity. We analyzed the preparation for the second trial in each sequence: cue period (10 timesteps), delay period (20 timesteps), and the initial trial period (10 timesteps). We fit SSMs across five levels of latent dimensionality (16, 40, 64, 88, 112), initializing the parameters with SSID at a horizon of 112.

SSM structure and preprocessing: EEG

We followed similar procedures for the Hall-McMaster and Arnau datasets as we did in the RNNs. We split the data into training and testing folds, preprocessing the inputs and observations in each separately. We epoched the recordings around the preparatory period, including the cue (200 ms), the delay period (Hall-McMaster: 400 ms; Arnau: 600 ms), and the first 200 ms of the trial (which we ensured was before any RTs through our trial rejection).

The SSM inputs included a (time-varying) bias, the current task, the previous task, the cue identity (with separate regressors contrasting the cue symbols within each task), cue repetitions vs. alternations for repeat trials, task switch vs. task repeat, the upcoming trial's reaction time, and the previous trial's reaction time. For the initial conditions, we included an intercept, the previous task, the previous reaction time, and the upcoming reaction time. We expanded the predictors with a set of cubic B-splines (Hall-McMaster: 20 bases per predictor, Arnau: 25 bases per predictor). Each predictor was z-scored across trials to standardize and reduce collinearity.

We preprocessed the EEG electrodes using PCA, projecting the voltage timeseries into PCs accounting for 99% of the variance. We used PCA because the observations were degenerate due to spatial proximity, independent components analysis, and bad-channel interpolation. PCA also reduced the computational demands of these analyses by reducing the observation dimensionality and making the observation covariance closer to diagonal. We estimated the PCs using only the training data, and then projected the test data onto these PCs.

As described above, the EM procedure cycled between estimating the latent states and updating the parameters. For EEG datasets, we fit the model across eight different levels of latent dimensionality: (16, 32, 48, 64, 80, 96, 112, 128). We set the maximum number of iterations to 20,000, measuring the test-set log-likelihood every 100 iterations. We terminated the fitting procedure if either the total data likelihood or the test likelihood stopped increasing.

SSID procedure

To provide the best initialization across the latent dimension hyper-parameter, we estimated SSID for a horizon and latent dimensionality that matched the largest tested model (128 latent factors). We then truncated these systems for smaller models (as latent states from CVA are ordered by their singular values). We found that SSID was enhanced by reshaping the trial-wise data into a long timeseries (e.g., allowing for larger numbers of factors and low frequencies). During SSID, we also only included inputs in the initial timesteps of each epoch, reducing the collinearity between lagged inputs. Note that we removed the test set before SSID (which was on separate experimental blocks to minimize temporal proximity). The EM procedure, unlike SSID, was not fit on concatenated epochs and used all of the inputs. In practice, we found that this SSID procedure provided a good initial guess of the parameters while avoiding requiring multiple initializations, and was especially important when there were more factors than observation dimensions.

Parameter recovery

We validated that the combination of model, data, and fitting procedure could produce identifiable parameter estimates under a known generative model. First, we used the SSM generative model to create a synthetic dataset based on a participant's parameters and inputs. We then estimated the parameters of this synthetic dataset using EM. Next, we aligned the estimated and recovered parameters, as these estimates are only identifiable up to an invertible transformation ($T = \text{Diagonal}(W); A' = T^{-1}AT; B' = T^{-1}B; C' = CT$; see [45] §2.5). Intuitively, we could shuffle the ordering of the latent factors without changing the likelihood. Finally, we correlated the generating and recovered parameters (correlating the lower triangle of the Cholesky factors for covariance matrices).

Our recovery was overall very high, but we had somewhat poorer recovery of input matrices. We found that this strongly depended on the norm of the input matrix column that we were trying to recover: some predictors were weakly encoded, and these were difficult to recover. It also likely reflects the correlation between proximal temporal bases.

Model comparison

Bayesian model selection

We selected the number of latent factors for subsequent analyses using a standard Bayesian model selection procedure.[93] This analysis estimates the expected probability of each model in the population. From these posterior mixture weights, we computed a 'protected exceedance probability,' the probability that a model is the most popular within the tested set of models, corrected for random guessing. While the best-fitting model in the Arnau dataset had 128 factors, we used the second-best model (112 factors) to allow for fair comparisons with Hall-McMaster and RNN models.

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

Sensor-level null models

To compare to a set of simpler null hypotheses, we fit a set of four models directly to the electrode PCs (i.e., without a latent embedding beyond the PCA). First was an intercept-only model (i.e., the standard R^2). Second was an encoding analysis using the same spline basis set as the SSMs (i.e., a linear additive model similar to EEG methods like the unfold toolbox, [94]). Third was a vector autoregressive (VAR(1)) model, estimating the multivariate relationship between sensor PC responses on adjacent timesteps. Fourth was a model that incorporated both autoregression and spline-basis encoding. Model four was the best-performing of these null models, so we used this as our benchmark.

For our SSM prediction metrics, we used the Cox-Snell 'generalized' R^2 [92], which uses the likelihood ratio between competing models rather than the ratio of their residual variance:

$$\text{generalized } R^2 = 1 - \left(\frac{\mathcal{L}_0}{\mathcal{L}_M}\right)^{2/n}$$

This metric is equivalent to the standard R^2 for Gaussian models with independently and identically distributed residual covariance,[95] but it has the advantage of reflecting our model's anisotropic, time-varying covariance from Kalman filtering.

EEG-optimized recurrent neural networks

To compare SSMs to a more complex, non-linear model, we predicted the observed EEG (component) timeseries using a new set of RNNs implemented in PyTorch. We fit eight sets of RNNs, each yoked to the same number of free parameters as SSMs at each of the different levels of latent dimensionality. The inputs to this RNN were the previous timestep's PC scores, and the same set of predictors as the SSM (formatted as constant inputs over each epoch, which fit better). These inputs were linearly projected into a set of hidden units, along with the previous hidden state, and then passed through a rectified linear unit (ReLU) activation function, using the standard PyTorch RNN implementation:

$$h_t = \text{ReLU}\left(x_{t-1}W_{ih}^{\top} + h_{t-1}W_{hh}^{\top} + b_{hh}\right)$$

$$h_{t=1} = x_0 W_0^{\top}$$

$$\hat{y} = h_t W_{ih[y]}^{\top}$$

Note that for consistency with PyTorch, we use h_t to refer to the latent state (SSM: x_t), and x_t to refer to the inputs (SSM: u_t). $W_{ih[y]}$ refers to the columns of W_{ih} used to encode the previous observations.

The PC scores on the next timestep were linearly decoded from the hidden state using the transpose of the observation embedding (a 'tied' parameterization; [96]). We found similar performance when we fit both the encoding and decoding layers, as this used more of our 'free parameter budget', trading off against the number of hidden units. The initial hidden state depended on the same inputs as the initial conditions of the SSM.

We fit 32 random initializations of this RNN to each participant and factor size. Each fitting session involved 5000 epochs of updating the parameters using backpropagation through time, using the mean square error of the next-timestep prediction as the loss function. Each batch contained all of the training data. Parameters were updated using the 'AdamW' learning rule (learning rate =.001, weight decay =.01; [79]). To evaluate each model's performance, we took the best test-set loss across all epochs and initializations.

Dynamic signatures of task control

Additive state decomposition

To isolate the dynamics of task representations, we leveraged the powerful superposition property of linear systems. In this case, SSMs can be factorized into an additive set of subsystems for each input, a procedure known as 'additive state decomposition'[97]. We used this decomposition to model 'task subsystems', in which state dynamics are only caused by task inputs. We modeled task inputs for switch trials using the difference between 'current task' and 'previous task' predictors, and task inputs for repeat trials using their sum, providing separate task subsystems for switch and repeat trials. Within each subsystem, the state evolves according to:

$$\boldsymbol{x}_t^{\text{task}} = A\boldsymbol{x}_{t-1}^{\text{task}} + B^{\text{task}}\boldsymbol{u}_{t-1}^{\text{task}}$$

$$\boldsymbol{x}_{t=1}^{\text{task}} = B_0^{\text{prevTask}}\boldsymbol{u}_0^{\text{prevTask}}$$

Since we contrast-coded the tasks (+1/-1), the dynamics for each task are symmetrical around the origin. For the initial conditions of this task subsystem, we used the 'previous task' component of the initial conditions ($B_0^{\text{PrevTask}}\boldsymbol{u}_0^{\text{PrevTask}}$), which is also symmetrical between switch and repeat trials for the same task.

To better equate task states across participants and datasets, we normalized the latent space using the univariate process noise ($T = \text{Diagonal}(W); A' = T^{-1}AT; B' = T^{-1}B; C' = CT$; see [45] §2.5).

Task energy

To quantify the influence of task representations, we used energy-based methods from Lyapunov analysis. This control theoretic tool is used to quantify asymptotic system properties for methods like controllability or observability analysis.[46,58,98] Controllability

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

quantifies the strength of the input-state coupling: how much of the task state is attributable to a given input. For a time-invariant linear dynamical system, the controllability Gramian ($\mathcal{G}_\infty$) defines the asymptotic state covariance that is attributable to a (white noise) input:

$$\mathcal{G}_\infty = A\mathcal{G}_\infty A^\top + BB^\top$$

$$= \sum_{t=0}^{\infty} A^t BB^\top A^{t\top}$$

$\mathcal{G}_\infty$ can be computed by solving the corresponding Lyapunov equation. To capture the influence of a particular sequence of time-varying inputs, we used the recursive Lyapunov equation to provide a time-resolved task Gramian[60]:

$$\mathcal{G}_t^{\text{task}} = A\mathcal{G}_{t-1}^{\text{task}}A^\top + \left(B^{\text{task}}\boldsymbol{u}_{t-1}^{\text{task}}\right)\left(B^{\text{task}}\boldsymbol{u}_{t-1}^{\text{task}}\right)^\top$$

$$\mathcal{G}_{t=1}^{\text{task}} = \left(B_0^{\text{prevTask}}\boldsymbol{u}_0^{\text{prevTask}}\right)\left(B_0^{\text{prevTask}}\boldsymbol{u}_0^{\text{prevTask}}\right)^\top$$

This task-dependent Gramian reflects how much the system has changed due to both immediate task inputs and the spread of task information through recurrence. It is analogous to how much of the system covariance is attributable to task inputs (except that the Gramian, unlike the covariance, is not mean-centered). Compare this Gramian update to how the covariance of a Gaussian distribution (underlined below) changes under stochastic linear dynamics:

$$x \sim \mathcal{N}(\mu, \Sigma)$$

$$\epsilon \sim \mathcal{N}(0, BB^\top)$$

$$y = Ax + \epsilon$$

$$y \sim \mathcal{N}(A\mu, \underline{A\Sigma A^\top + BB^\top})$$

We summarized task energy using a standard metric of 'average controllability'[46], taking the trace of $\mathcal{G}$ at each timestep. This metric is similar to the 2-norm of the task state,[34] but it does not include temporal cross-terms. We found similar results using the 2-norm of the task state (Switch- Repeat contrast: Hall-McMaster: $d = 0.88, p = 4.4 \times 10^{-5}$; Arnau: $d = 0.55, p = 0.01$). We log-transformed the average task energy to reduce skew, and then contrasted the Switch and Repeat conditions.

Switch similarity

To characterize the similarity between task trajectories on switch and repeat trials, we measured the cross-lagged similarity between inferred task states for switch and repeat trials, similar to a temporal generalization analysis.[62,99] In practice, we computed the cosine of the angle between task states across switch and repeat conditions at different cross-lags (i.e., computing the similarity between every timestep in the repeat subsystem to every timestep in the switch subsystem). We tested this similarity against zero at the group level using TFCE (STAR Methods).

Initial centrality

To understand whether task states converged to a common state during the ITI, we first measured the similarity between RNN hidden unit activations during the ITI. We averaged the hidden states during the ITI following each task, and then computed the (log) Euclidean distance between these average trajectories (i.e., the distances between the post-A states and the post-B states). Smaller values corresponded to states becoming more similar, and the slope of the line provided the (exponential) rate of convergence.

To understand whether RNNs and EEG converged to a 'task neutral' state during the ITI, we tested whether the terminal ITI state (i.e., at cue onset, the initial conditions of the SSM) was close to the midpoint of the states that were occupied during the trial. We developed a 'midpoint score' that quantified the relative Euclidean distance between the initial task state and each task state along the trajectory, which we computed for switch and repeat trials:

$$\delta_t = \frac{\|\boldsymbol{x}_1^A - \boldsymbol{x}_t^A\|_2 - \|\boldsymbol{x}_1^A - \boldsymbol{x}_t^B\|_2}{\|\boldsymbol{x}_1^A - \boldsymbol{x}_t^A\|_2 + \|\boldsymbol{x}_1^A - \boldsymbol{x}_t^B\|_2}$$

Initial interference

We quantified 'initial interference' in our SSM models as the strength of 'previous task' encoding at the initial timestep. The initial conditions included regressors for the previous task, so we used a norm of these previous-task inputs ($\text{logrms}\left(B_0^{\text{prevTask}}\boldsymbol{u}_0^{\text{prevTask}}\right)$).

To quantify the differences in interference between the two EEG experiments, we computed the ratio of the median interference value between studies (i.e., the difference between the log rms). We found that this rescaling closely matched the Arnau dataset to the Hall-McMaster dataset. We used this scaling factor to rescale B_0^{prevTask} for the Arnau dataset, re-computing the task trajectories and switch similarities. We used the same bootstrap test to compare the switch similarity between the two EEG datasets as we used to compare the RNN and EEG datasets.

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

We confirmed that the enhanced alignment between datasets was not a trivial byproduct of rescaling the parameters to make them more similar. We computed the initial encoding of the current task ($\text{logrms}(B^{\text{task}}\boldsymbol{u}_1^{\text{task}})$), found the ratio of the medians between the experiments, rescaled B^{task} in the Arnau dataset, and then recomputed the task trajectories and switch similarities.

We repeated this yoking procedure for RNNs with long and short ITIs, finding that matching initial interference produced similar patterns of switch similarity.

Corrections for multiple comparisons

To correct for multiple comparisons over time while accounting for temporal autocorrelation, we used threshold-free cluster enhancement throughout (TFCE; [100]) with a set of EEG-optimized parameters (H=2, C=1; from[101]; 1000 permutations for temporal generalization and 10,000 permutations for traces).

EEG-RNN comparison

To evaluate the similarity between dynamic signatures between RNNs and humans, we first subsampled human task trajectories (100 or 125 timesteps) to match the duration of networks' task trajectories (40 timesteps).

We averaged the outcome measures (Switch Similarity, Initial Centrality, or Task Energy) within RNNs and EEG, and then computed the similarity between RNNs and EEG. For Switch Similarity and Task Energy, we computed the similarity using the 'congruence coefficient', taking the cosine of the vectorized measures.

For Initial Centrality, we cared about the proximity to zero for a measure that was already normalized. For these reasons, we used an R^2 similarity between EEG and RNNs:

$$R^2(x,y) = 1 - \frac{\sum (x-y)^2}{\sqrt{(\sum (\bar{x}-x)^2)(\sum (\bar{y}-y)^2)}}$$

This metric gives a similarity value that ranges from $-\infty$ to 1. This metric can be negative if the distance between modalities is greater than the variance within modalities (recall that the modalities are not fit to each other).

To produce a bootstrap estimate of these metrics, we sampled with replacement 10,000 times within the RNN and EEG groups (e.g., drawing 512 RNNs and 30 EEG participants at random, with replacement). Within each bootstrap sample, we averaged our measures and computed our metrics as outlined above. We then computed the 2.5th and 97.5th percentiles of this bootstrap distribution.

To get an estimate of the difference between switch-trained RNNs, we sampled with replacement from (1) 1-Trial RNNs, (2) 2-Trial RNNs, and (3) EEG participants. We computed similarity metrics between 1-Trial RNNs and EEG, and between 2-Trial RNNs and EEG, and then we took the difference between these metrics. This approach provided a distribution of similarity differences, compared within the same bootstrap sample of EEG participants. We expect EEG variability to be higher than RNNs due to noisy measurements and a smaller sample, and by accounting for EEG variability in this difference distribution, our difference measure could be more sensitive than looking for overlap in the confidence intervals for each metric.

EEG task state analyses

EEG task trajectory visualization

Using the decomposition and symmetry principles afforded by the linear system, we visualized the estimated latent trajectories of task representations between switch and repeat trials using group-level singular value decomposition (SVD).

To visualize the latent trajectories over time, we spatially concatenated participants' switch and repeat task trajectories into a wide 2D 'temporal' matrix $((\text{switch timesteps} + \text{repeat timesteps}) \times (\text{factors} \times \text{participants}))$. We then used SVD to extract separate scores for the switch and repeat subsystems. Since we only have two tasks, tasks were contrast coded (-1/1), which meant that the trajectories were symmetrical around an interpretable zero point (i.e., no task inputs). Despite this redundancy, we plotted the trajectories for all four conditions (i.e., task $\times$ switch) for illustrative purposes.

EEG performance alignment

To confirm that the inferred task states were relevant for task performance, we tested whether trial-level brain states predicted upcoming reaction times.

First, we projected each participant's task trajectory for switch and repeat subsystems back into the observation space ($\hat{\boldsymbol{y}}_t^{task} = C\boldsymbol{x}_t^{\text{task}}$). We computed the trial-averaged cosine similarity between the task state and the EEG response ($\text{mean}(\cos(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t^{task}))$). We included this 'task score' in a linear mixed-effects regression predicting log-transformed reaction times (MATLAB's fitlme), alongside predictors accounting for the intercept, task, switch vs. repeat, and the task $\times$ switch interaction. We included maximal random intercepts and slopes, centered predictors within each participant, and confirmed that design matrices passed collinearity diagnostic tests. We tested whether group-level 'task score' estimates were significantly different from zero after the Satterthwaite correction for degrees of freedom. We found similar results when also controlling for the trial-averaged norm of the EEG and task state.

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

Supplementary analyses

Procrustes alignment

To align the RNN task states to EEG, we first concatenated the switch and repeat task trajectories, after temporally downsampling the EEG datasets to match the RNNs. Our bootstrapping procedure mirrored our other tests of RNN-EEG alignment. We drew random RNNs and EEGs with replacement, averaged their (concatenated) task trajectories, and then computed the Procrustes distance between these trajectories using the MATLAB procrustes function. Refer to Figure S4.

Briefly, Procrustes alignment finds the optimal rotation, translation, and scalar stretch (i.e., the same factor applied to all dimensions) to minimize the mean squared error between two matrices. This scalar stretch ('similarity transformation') preserves the geometry of the data and reduces overfitting. The MATLAB function returns a metric between 0 and 1, which we aggregated across bootstrap samples. We also computed the difference between distances within each sample, providing a direct contrast within the same bootstrap sample of EEG participants. We computed the joint p-value by computing the proportion of bootstrap samples less than 0 for each EEG study, and then aggregating these p-values with a z-test (converting the p-values to z-values using the inverse cumulative normal distribution, averaging, and then converting this average z back to a p-value; [102]).

Variable-ITI RNNs

We yoked the ITI range to match long-ITI RNNs to the Hall-McMaster dataset, and short-ITI RNNs to the Arnau dataset. We trained RNNs with an equal number of short, medium, and long ITIs, using different amounts of slack time at the end of each sequence. All other fitting and quantification procedures were the same as the main experiment. Refer to Figure S5.

Response-penalized RNNs

To produce 1-Trial models with response convergence, we modified the RNN training target. Now, after the first trial, 1-Trial RNNs had an ITI without stimulus inputs (with the ITI matching their trial-spacing condition). During this ITI, we had a target output of zero for both response units, training RNNs not to produce outputs during this period. We only trained 1-Trial RNNs with this procedure, as they represent the alternative hypothesis (i.e., that convergence may reflect response dynamics rather than task dynamics). All other fitting and analyses were the same. We just plotted the initial centrality for these models, since this was a decisive divergence from the EEG findings. However, 2-Trial RNNs were better matched to the EEG datasets across the board. Refer to Figures S6A–S6C.

Activation convergence

We tested whether the convergence of RNNs into a common state during the ITI was due to relaxation towards the origin. To measure convergence towards the origin during the ITI ('activation convergence'), we took the average of the hidden unit activation between the trajectories following each task, and measured the Euclidean distance from this coordinate to the origin. If the RNNs were converging towards the origin during the ITI, then the distance between their post-task states should decrease at around the same rate as their average distance from the origin. This did not appear to be the case: RNNs converged towards a common post-task state, but showed little movement towards the origin. Refer to Figure S6.

Contraction analysis

We formally tested whether RNNs converge into a common state during the ITI due to attractor dynamics by using nonlinear contraction analysis.[103,104] This approach quantifies whether pairs of trajectories converge to a common state over time.

We may have one trajectory during the ITI after task A (from $\boldsymbol{x}^A_{t+1} = f_\theta(\boldsymbol{x}^A_t)$) and another trajectory during the ITI after task B (from $\boldsymbol{x}^B_{t+1} = f_\theta(\boldsymbol{x}^B_t)$). We want to determine whether $\| \boldsymbol{x}^A - \boldsymbol{x}^B \|$ will converge to zero. To estimate this convergence, we analyze the dynamics of the displacement between these two trajectories. If these trajectories were infinitesimally close, this change in displacement is equal to the Jacobian of $f_\theta(\boldsymbol{x})$ ($\mathrm{Jac}(f_\theta(\boldsymbol{x}))$), which is the matrix of first-order partial derivatives of the function f_θ with respect to $\mathbf{x}$. Since these trajectories are not infinitesimally close, we integrate the Jacobian over the line that connects the two trajectories at each timestep.

To characterize whether the displacement dynamics are contracting, we can analyze the dynamics over the modes of the displacement Jacobian. Here, we used the absolute eigenvalues of the Jacobian ($|\lambda(\mathrm{Jac}(f_\theta(\boldsymbol{x})))|$) to capture the long-term behavior of this system (e.g., ignoring transient amplification from non-normal dynamics). The maximum eigenvalue will give a bound on the system's asymptotic contraction rate, and the eigenspectrum will reveal the range of contraction rates. Thus, our worst-case estimate of whether a given pair of trajectories is contracting at a given timestep is:

$$\int_0^1 \left| \lambda^{max}\left(\mathrm{Jac}\left(f_\theta\left(w x^A_T + (1-w) x^B_T\right)\right)\right) \right| dw$$

In 128 RNNs per condition, we sampled 128 pairs of trajectories (pairs of post-A and post-B ITIs). We approximated the line integral between trajectories by averaging over 11 equally-spaced points. We report both the worst-case contraction (maximum eigenvalues) and the eigenvalue averaged within quintiles of the eigenspectrum. Refer to Figure S6.

Biased transition training

To explore the influence of biased transitions on RNN strategies, we trained 2-Trial RNNs that, during double-trial training, experienced 50% switch trials, 75% switch trials (mostly switched), or 25% switch trials (mostly repeated). The ITI convergence was quantified as in the main text.

The midpoint score was computed on the RNN hidden units (rather than on the SSM factors) and over the ITI (rather than during the cue interval). We computed the distance from the hidden state during the ITI to the state at the onset of the second trial's decision epoch (i.e., 10 timesteps after the stimulus onset, when the RNN training target began). We computed the same relative distance

Please cite this article in press as: Ritz et al., Inter-trial convergence of neural task states supports cognitive flexibility, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.07.028

function as in the main text (difference divide by sum), comparing the ITI states after each task to the 'decision state' corresponding to that same task. Negative values indicate ITI states closer to the decision state for a repeating task (repeat bias), zero indicates the task-neutral state, and positive values indicate ITI states closer to the decision state for the alternating task (switch bias). Refer to Figures S6H–S6J.

Gate ablations

To assess the role of the reset and update gates for RNNs' performance, we trained a series of models with ablated or frozen gates (Figures S7A–S7C).

The ablated models consisted of (1) fixing the reset gate to be open (full use of previous hidden states), (2) fixing the update gate to be open (full use of new state), or (3) training an Elman RNN[105] (no gating). These parameters were ablated for the entirety of training. We trained GRU-RNNs with a tanh activation function and learning rate of 0.01, whereas the ElmanRNN was trained with a ReLU activation function and a learning rate of 0.001 (which improved its performance). To assess the long-term influence of gating, we trained these models for 450 epochs of one-trial sequences, followed by 450 epochs of two-trial sequences, all with short ITIs. Free parameters were matched across networks (108 hidden units for the full model, 134 units for ablated GRUs, 190 units for Elman RNN).

For the frozen models, we first trained full GRU-RNNs for 450 epochs on one-trial sequences (as in the original experiment). We then froze the weights and biases for (1) the reset gate, (2) the update gate, or (3) both gates, while training all other parameters for 450 epochs of the two-trial sequences. We compared the peak performance of frozen and ablated RNNs by comparing each RNNs' best test loss during the two-trial epochs against the full models' performance.

Gate-free RNNs

We trained no-gate RNNs using the same procedure as our ablation experiments, but here for the same 450–50 training scheme as GRU-RNNs. Several no-gate RNNs models timed out during the SSM fitting procedure, leaving 342 short-ITI RNNs and 311 long-ITI RNNs. Given the stark differences between RNNs and EEG, and the secondary importance of this analysis, we considered this sample size to be sufficient. Refer to Figures S7D and S7E.

Software

1.Programming languages

(a) Julia (v1.11.2)
(b) Python (v3.10.13)
(c) Matlab (R2023b, R2025b)

2.Software packages

(a) PyTorch (v2.4.0,[75]) www.pytorch.org
(b) MatlabTFCE

www.github.com/markallenthornton/MatlabTFCE

(c) ControlSystemIdentification.jl (v2.10.2) www.github.com/baggepinnen/ControlSystemIdentification.jl
(d) FieldTrip[76]
(e) EEGLAB (v2024.2;[77])

3.Visualization

(a) Scientific color maps[78]

The author(s) are pleased to acknowledge that the work reported on in this paper was substantially performed using the Princeton Research Computing resources at Princeton University, which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology's Research Computing.

Current Biology, Volume *36*

Supplemental Information

Inter-trial convergence of neural task states supports cognitive flexibility

Harrison Ritz, Aditi Jha, Nathaniel D. Daw, and Jonathan D. Cohen

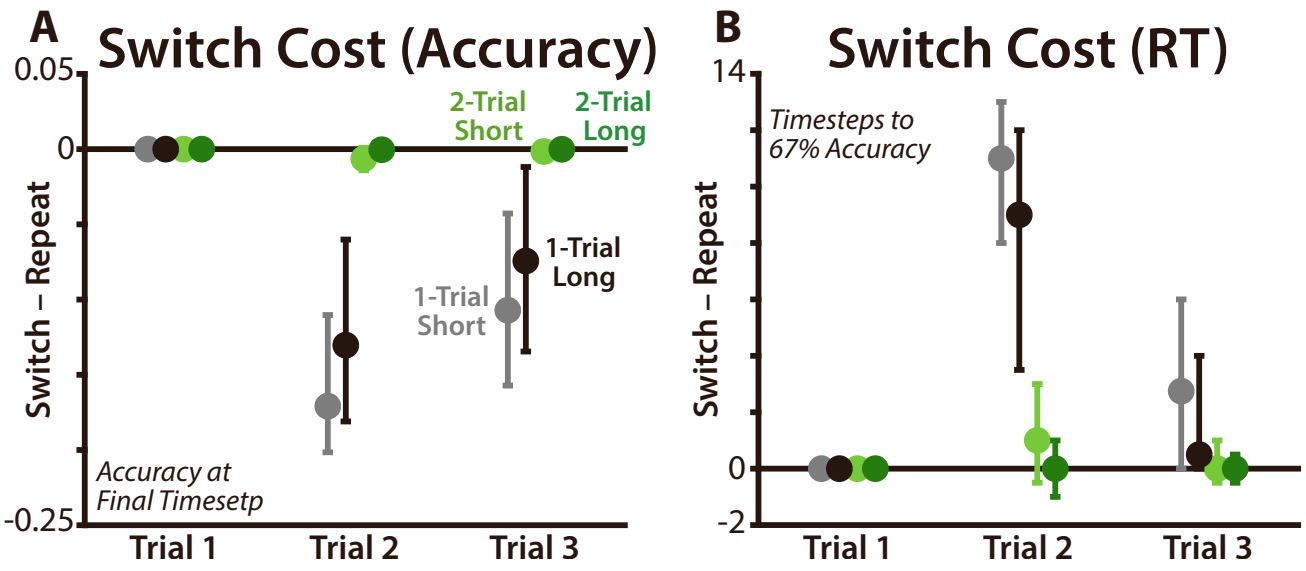


Figure S1: Performance switch costs across RNNs. Related to Figure 1.
(A) Difference in RNNs' final Accuracy between switch and repeat trials, plotted for each RNN condition on 3-trial sequences. Third-trial switch effects are following a repeat. Error bars reflect quartiles of the between-RNN distribution.
(B) Difference in in RNNs' RT between switch and repeat trials, plotted for each RNN condition on 3-trial sequences. Third-trial switch effects are following a repeat. Error bars reflect quartiles of the between-RNN distribution.

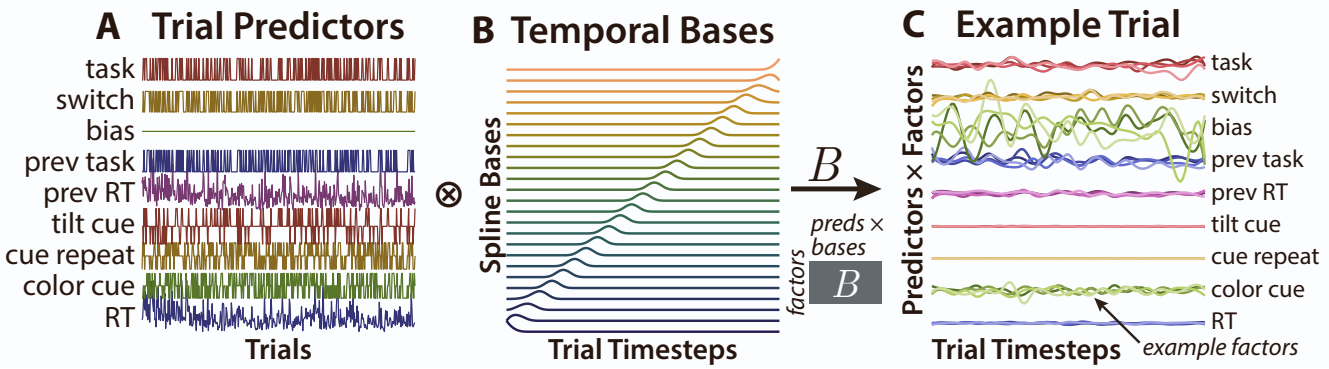


Figure S2: Input spline basis. Related to Figure 2.
(A) Across-trial predictors from the SSMs fit to EEG.
(B) Temporal B-spline bases. These bases tiled the trial epoch, allowing us to estimate smoothly time-varying inputs. SSM inputs were the Kronecker product between across-trial predictors and within-trial bases (i.e., every combination of predictor and basis).
(C) Estimated instantaneous input effects for an example trial. Input effects are estimated using the dot product between the inputs and the fitted input matrix ('B' matrix, mapping inputs to factors). Shaded lines for each predictor indicate input effects in four example factors. Note that these are 'instantaneous' input effects because inputs also spread through the recurrence matrix (not shown here).

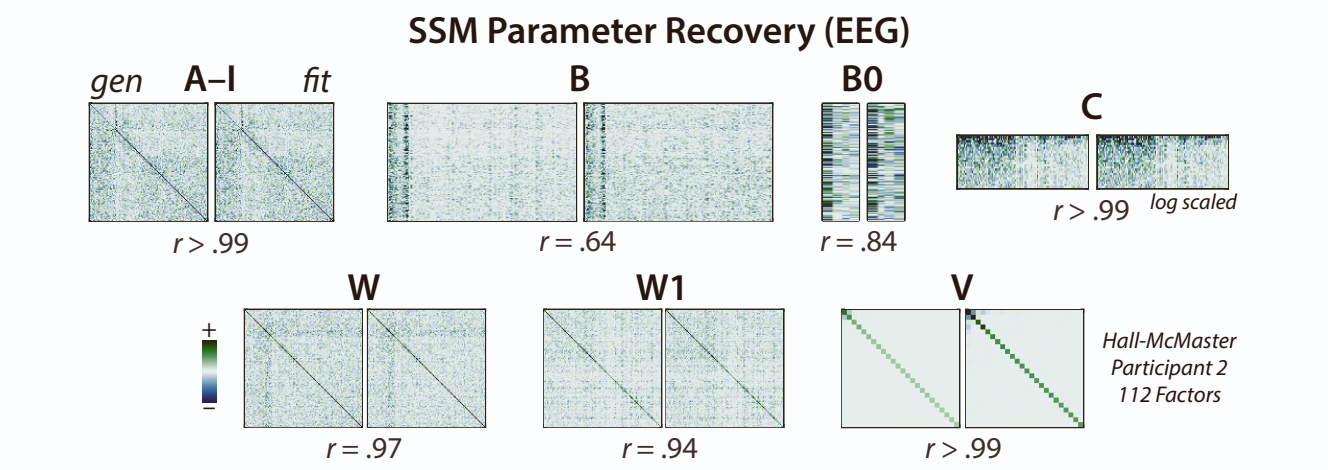


Figure S3: SSM parameter recovery in EEG. Related to Figure 3.
Generating and fit parameters for an example participant in Hall-McMaster. Parameters were linearly aligned to account for degeneracy in the latent system (see STAR Methods).

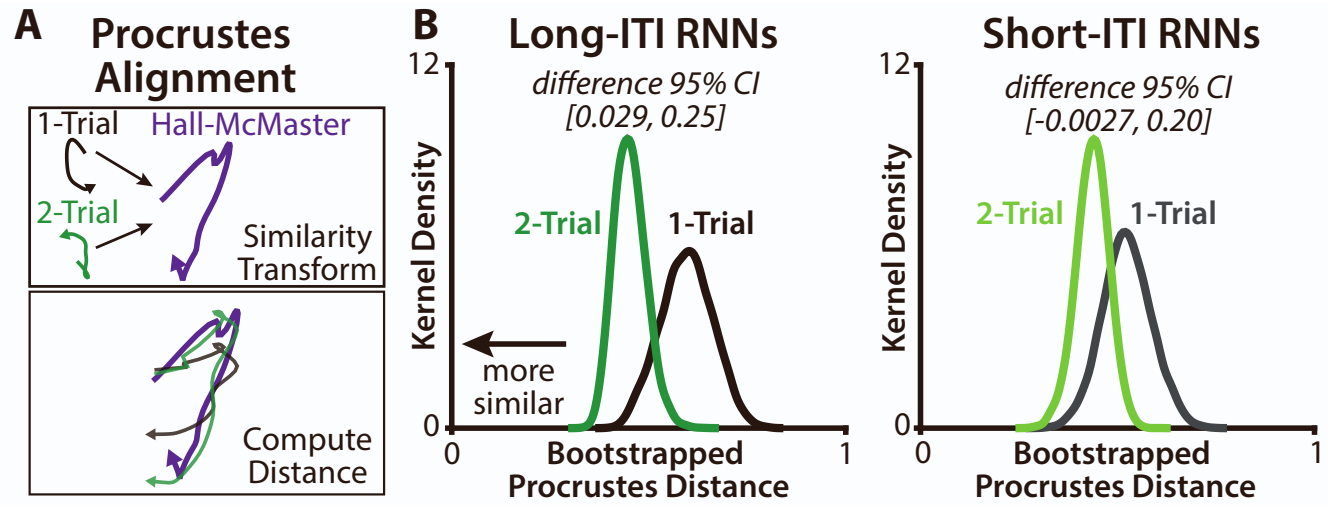


Figure S4: Procrustes alignment between RNNs and EEG. Related to Figure 4.
(A) We used procrustes alignment to apply a 'similarity transform' between RNN and EEG task trajectories. This similarity transform applied rotations, translations, and isotropic scaling (i.e., all dimensions are scaled by the same factor) to minimize the mean squared error between RNNs and EEG. The post-alignment distance metric is bound between 0 and 1. While SVD projections for a single trajectory are shown here for illustrative purposes, procrustes distances were computed using all factors, on the temporally-concatenated switch and repeat trajectories. **(B)** 2-Trial RNNs were more similar to EEG than 1-Trial RNNs. Joint bootstrapped *p*-value: 4.2×10^{-5}. For more details, see STAR Methods.

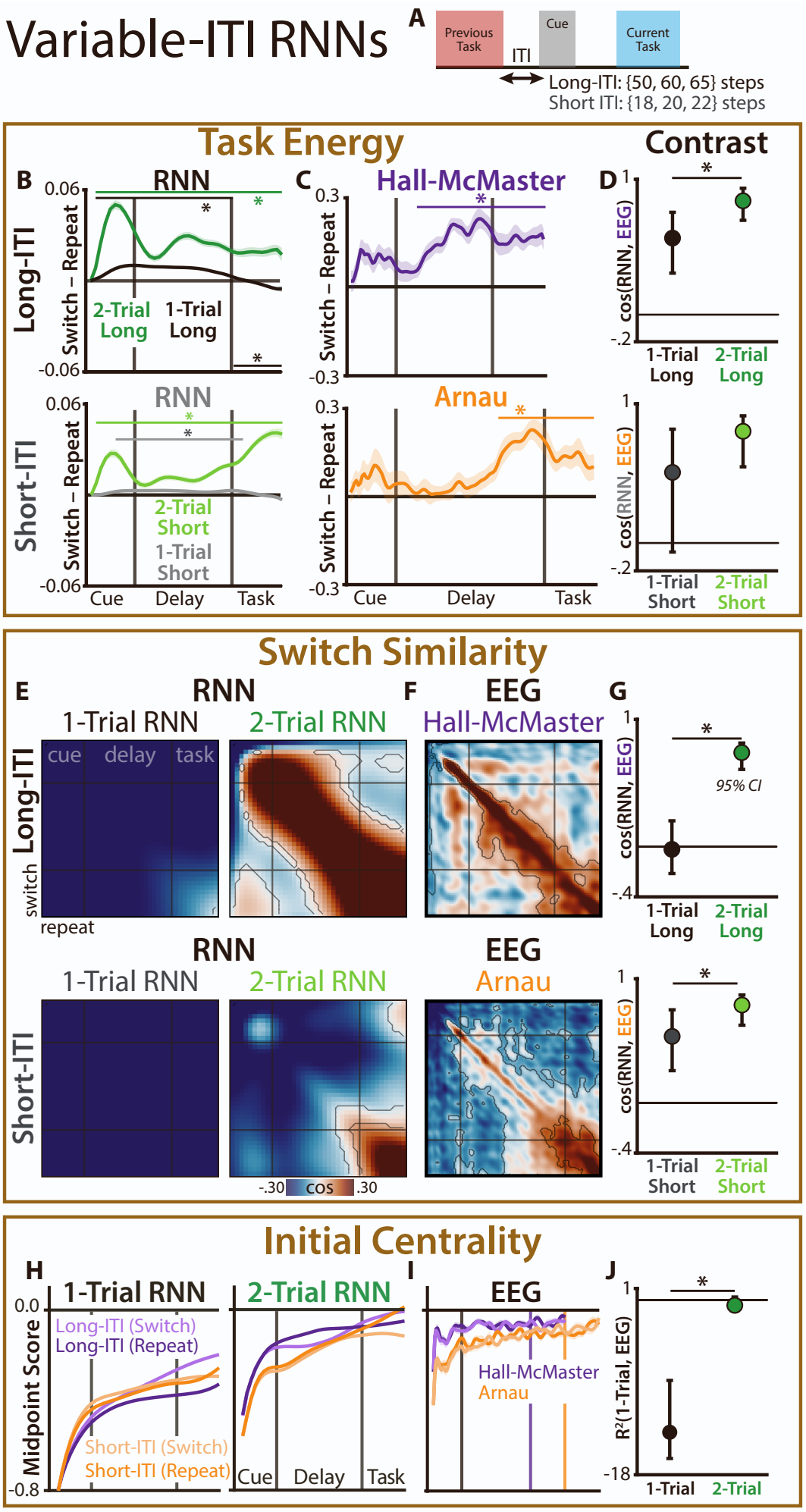
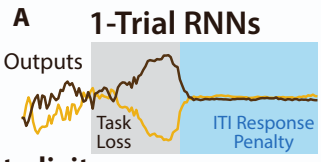

Variable-ITI RNNs
A
Previous Task
ITI
Cue
Current Task
Long-ITI: {50, 60, 65} steps
Short ITI: {18, 20, 22} steps
Task Energy
Contrast
B
RNN
Long-ITI
Switch – Repeat
0.06
-0.06
2-Trial Long
1-Trial Long
C
Hall-McMaster
0.3
-0.3
D
cos(RNN, EEG)
1
-.2
1-Trial Long
2-Trial Long
RNN
Short-ITI
2-Trial Short
1-Trial Short
Arnau
Cue
Delay
Task
1-Trial Short
2-Trial Short
Switch Similarity
E
RNN
1-Trial RNN
2-Trial RNN
F
EEG
Hall-McMaster
G
cue
delay
task
switch
repeat
95% CI
-.4
Arnau
-.30
cos
.30
Initial Centrality
H
1-Trial RNN
2-Trial RNN
I
EEG
J
Midpoint Score
0.0
-0.8
Long-ITI (Switch)
Long-ITI (Repeat)
Short-ITI (Switch)
Short-ITI (Repeat)
Hall-McMaster
Arnau
R²(1-Trial, EEG)
-18
1-Trial
2-Trial

Figure S5: RNNs with variable ITIs reproduce core findings. Related to Figure 5.
(A) RNNs were trained with variable inter-trial intervals (ITIs). ITIs took one of three values across sequences, yoked to the variabily in the EEG experiments. Each training epoch had an equal number of sequences with each ITI. Compare the panels below to Figure 5, Figure 6, and Figure 7.
(B-C) Task Energy: difference in task energy between switch and repeat trials, for (B) RNNs and (C) EEG. Asterisks indicate $p < .05$ (TFCE-corrected).
(D) Bootstrapped cosine similarity between EEG and 1-Trial vs. 2-Trial RNNs.
(E-F) Switch Similarity: Cross-lagged similarity between SSM task states on switch vs. repeat trial, for (E) RNNs and (F) EEG. Cardinal lines indicate boundaries between cue, delay, and task events. Contours indicate $p < .05$ (TFCE-corrected).
(G) Bootstrapped cosine similarity between EEG and 1-Trial vs. 2-Trial RNNs.
(H-I) Initial Centrality: relative distance between SSM initial conditions and each task state, for (H) RNNs and (I) EEG. Zero indicates initial conditions are at the midpoint.
(J) Bootstrapped R^2 between EEG and 1-Trial vs. 2-Trial RNNs. For more details, see STAR Methods.

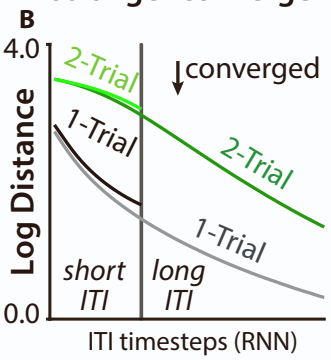
Response Penalty

1-Trial RNNs: ITI Response Penalty

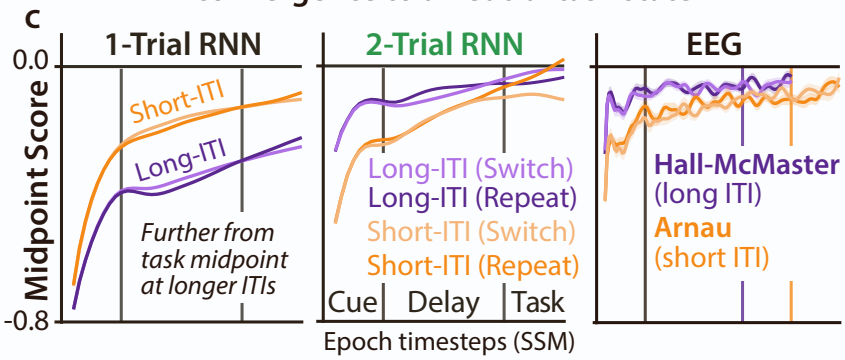
2-Trial RNNs: No Penalty (Original)



Response penalty elicits stronger convergence

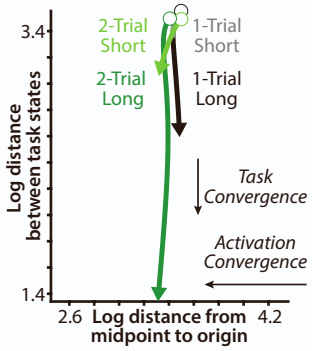


Response penalty does not elicit convergence to a neutral task state

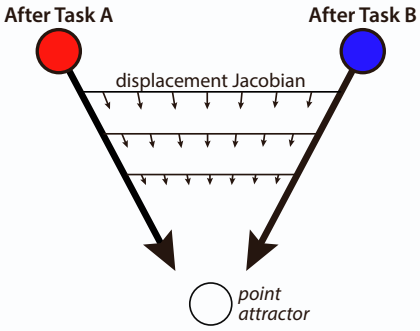


Contraction Analysis

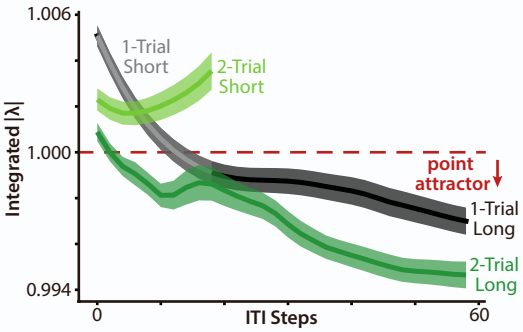
D Task vs. Activation Convergence



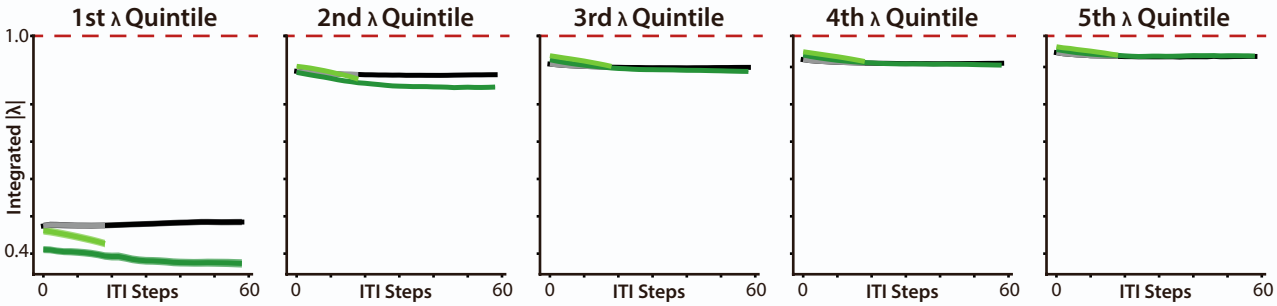
E Contraction Analysis



F Worst-Case Contraction

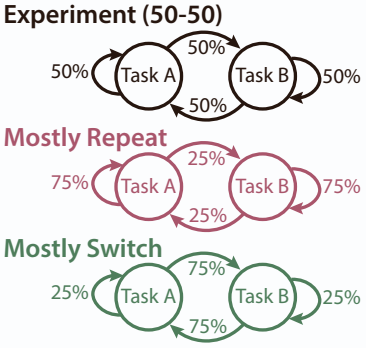


G Contraction Spectrum

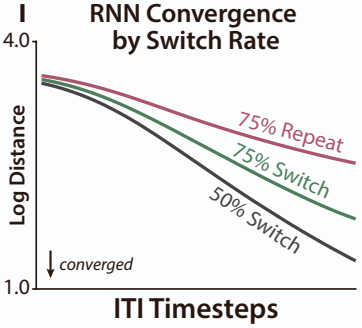


Switch Probability

H Experiment (50-50)



I RNN Convergence by Switch Rate



J RNN Centrality by Switch Rate

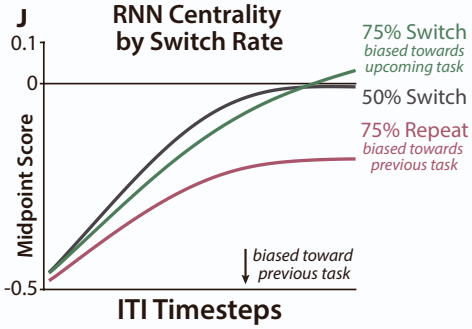


Figure S6: Validation of ITI Convergence. Related to Figure 7.
(A) 1-Trial Response-Penalized RNNs were similar to 1-Trial RNNs the main experiment, but also encountered a post-trial ITI during which there was a response penalty. These RNNs were compared to the 2-Trial RNNs from the main experiment.
(B) Response-Penalized RNNs showed stronger convergence to a Neutral state than standard 2-Trial RNNs. Compare to Figure 4F.
(C) Response-Penalized RNNs produce the opposite pattern of Midpoint scores from 2-Trial RNNs and EEG. Unlike EEG, Response-Penalized RNNs were further from the task midpoint at longer ITIs. Compare to Figure 7. For more details, see STAR Methods.
(D) The distance between the RNN hidden states after each task (y-axis) converged much faster than their mean distance from the origin (x-axis). This is consistent with RNNs converging into a task-specific neutral state rather than a global low-energy state. For more details, see STAR Methods.
(E) To characterize the rate of convergence in RNNs during the ITI, we used a nonlinear 'contraction analyses'. Contraction analyses characterize the convergence of pairs of trajectories towards a common fixed point through an analysis of the Jacobian of their displacement. At each timestep, we computed the Jacobian of the displacement between pairs by integrating the absolute eigenvalues of the Jacobian over the line connecting the pair. Eigenvalues less than 1 will asymptotically decay to zero. For more details, see STAR Methods.
(F) The maximum integrated eigenvalue reveals whether the whole system is contracting towards a fixed point. No RNNs showed strong expansion (largest max eigenvalue less than 1.006). 2-Trial Long-ITI networks tended to have faster global contraction than 1-Trial RNNs.
(G) Quintiles of the integrated eigenvalues. Most eigenvalues were less than 1, indicating that the system is primarily stable. 2-Trial RNNs exhibited stronger contraction for the fastest contracting eigenvalues (i.e., 1st Quintile), suggesting that neutral state convergence may occur within a subspace of the overall system.
(H) 2-Trial Long-ITI RNNs had 2-trial training schedules with a 50% switch rate (main analysis), a 75% repeat rate ('Mostly Repeat'), a 75% switch rate ('Mostly Switch'). All analyses are performed on the RNN hidden unit activity.
(I) RNNs with 50% switch rate had the strongest convergence into a common state (distance between ITI states following each task), and RNNs with 75% switch rate had the least convergence into a common state.
(J) The task midpoint was defined as the midpoint between the hidden states at the decision onset for each task. The midpoint score function was similar to the main text (normalized difference), but here computed on RNN hidden states during the ITI. RNNs with a 50% switch rate converged into a task-neutral state during the ITI, consistent with the main text. RNNs with a 75% repeat rate stayed closer to the previous task. RNNs with a 75% switch rate moved towards the alternating task. For more details, see STAR Methods.

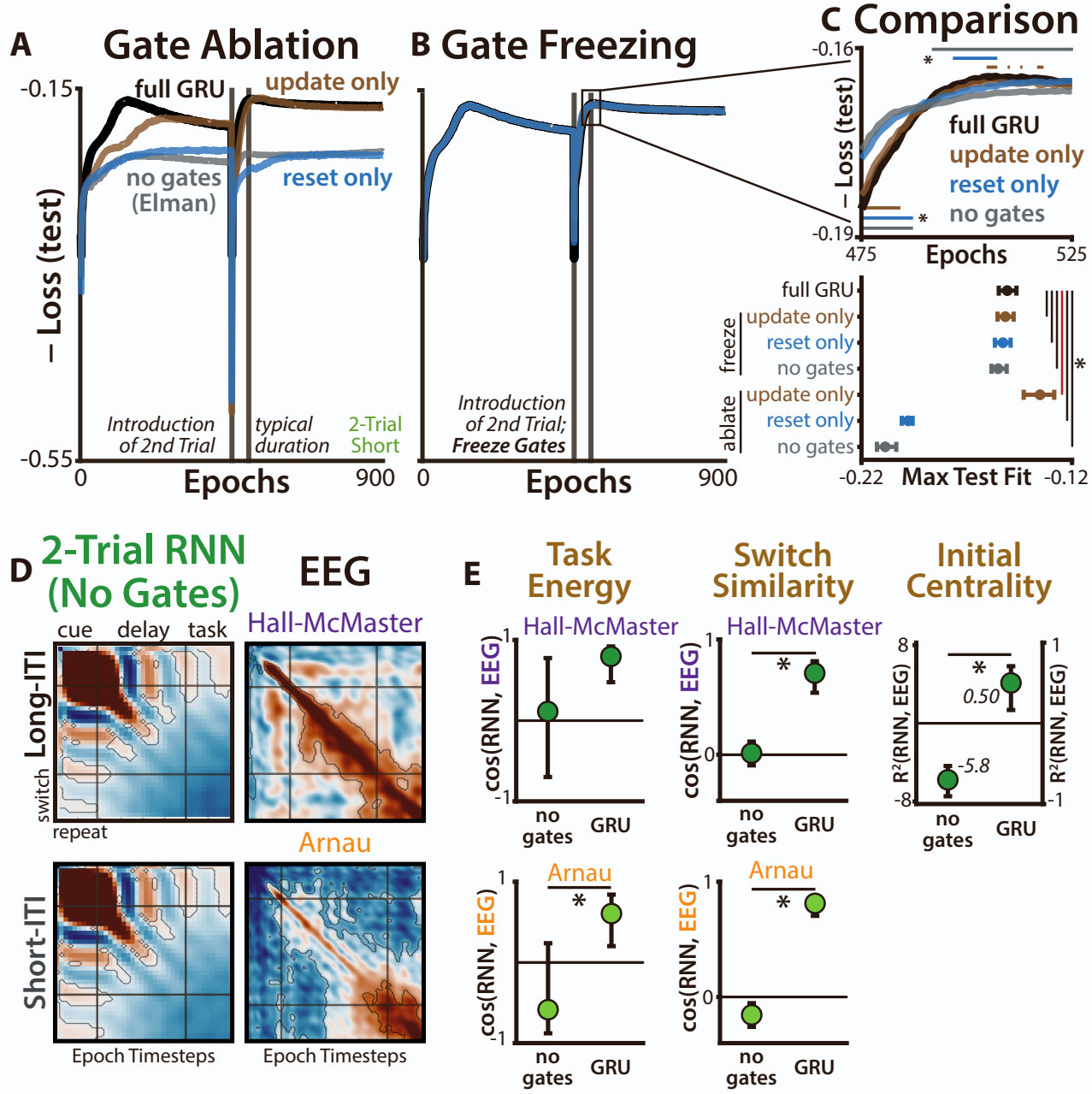


Figure S7: RNN gating mechanisms improve task performance and EEG correspondence. Related to Figures 7.

(A-B) Test performance for 2-Trial RNNs that had either ablated gates (A; open for the entire training period) or frozen gates (B; gates no longer trained after single-trial epochs). Shown for short-ITI test sequences (no input noise). Horizontal lines indicate the onset of double-trial training (epoch 450) and the standard training duration for the main simulation (epoch 500). For more details, see STAR Methods.

(C) Top: zoomed-in view of the frozen gate loss at peak performance. Asterisks and lines indicate $p < .05$, corrected using threshold-free cluster enhancement (TFCE; STAR Methods). Bottom: Pairwise comparisons between peak test performance in the Full GRU model against the frozen and ablated models. Asterisks indicate significant differences (bootstrap test), error bars indicate quartiles.

(D) Cross-lagged similarity between SSM task states on switch vs. repeat trial for 2-Trial No-Gate RNNs (left; i.e., 'Elman' RNNs) and EEG (right). Compare to Figure 6A-B. Contours indicate $p < .05$ (TFCE-corrected).

(E) Comparison between No-Gate RNNs and the 2-Trial GRU RNNs from the main analysis in terms of how well they match EEG. Compare to Figure 5, Figure 6, and Figure 7. Error bars indicate bootstrapped confidence intervals; asterisks indicate $p < .05$ from a bootstrap test. For more details, see STAR Methods.