# Dynamic neural control of task representations in humans and neural networks

**Harrison Ritz, Aditi Jha, Jonathan Pillow, Nathaniel Daw,** and **Jonathan D. Cohen**
Princeton Neuroscience Institute, Princeton University
contact: hritz@princeton.edu

## Abstract

**Cognitive psychologists often model task preparation using dynamical systems theory, however the neural correlates of these cognitive dynamics remain poorly understood. We bridged between cognitive and neural theories by fitting linear dynamical systems to human EEG recordings during task switching. Using a control theoretic analysis of the fitted dynamical system, we found that participants showed stronger propagation of task information when switching tasks than when repeating tasks. Similar signatures of task control were evident in task-optimized neural networks, consistent with this neural marker of task reconfiguration reflecting an optimality principle.**



**Keywords: Task Switching, EEG, Dynamical System, Latent State Space, Recurrent Neural Network**


**Introduction** Exciting recent work in cognitive psychology models task preparation as a dynamical system (Musslick, Jang, Shvartsman, Shenhav, & Cohen, 2018; Jaffe, Poldrack, Schafer, & Bissett, 2023), however the neural bases of these cognitive dynamics remain poorly understood. Here, our research goals were to 1) validate methods for fitting macroscale state space models (SSMs) to human EEG recordings, 2) quantify how people control neural dynamics to implement task-relevant brain states, and 3) compare metrics of neural control across humans and recurrent neural networks.

**Task & Sample.** A complete description is available in Hall-McMaster, Muhle-Karbe, Myers, and Stokes (2019). Thirty human participants performed a cued task switching experiment during 61-channel scalp EEG recording (Fig 1A). On each trial, participants responded to a compound stimulus based on either its color or its shape. Before each trial, participants were cued to the task-relevant feature and whether they would earn bonus payment for good performance (not shown). We analyzed epochs without recording artifacts or errors on the previous/upcoming trial ($M = 463$ trials).

**Model fitting.** State space models are statistical models of neural population activity with growing popularity in computational neuroscience (Smith & Brown, 2003; Macke et al., 2011; Linderman, Nichols, Blei, Zimmer, & Paninski, 2019). Here, we inferred how $\mathbf{y}_t$, the vector of EEG electrode voltages at time $t$, arises from the linear projection of the latent state vector $\mathbf{x}_t$ (i.e., neural generators; Fig 2A). Formally, $\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t$, where C is a matrix of projection weights and $\mathbf{v}_t \sim \mathcal{N}(0, R)$ is Gaussian noise. The latent state $\mathbf{x}_t$ evolves linearly over time according to $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{u}_{t-1} + \mathbf{w}_t$, where $A$ is the recurrent dynamics, $B$ is the projection of input vector $\mathbf{u}_t$ (e.g., task conditions), and $\mathbf{w}_t \sim \mathcal{N}(0, Q)$ is Gaussian noise.

We modeled inputs as boxcar functions over the task cue (Fig 2B). Inputs included a constant bias, task identity, the specific cue stimulus, task switch vs. repeat, subsequent reaction time (RT), high vs. low reward, and the two-way interactions between task and each of switch, RT, and reward.

We fit SSMs to each participant using a custom expectation-maximization procedure in Julia. Before fitting, we projected each participants' electrode timeseries onto principal components capturing 99% of the variance (15-28 components). We initialized our parameter estimates with subspace identification (Stone, Sagiv, Park, & Pillow, 2023), using an 'instrumental variable' method from `ControlSystemIdentification.jl`.

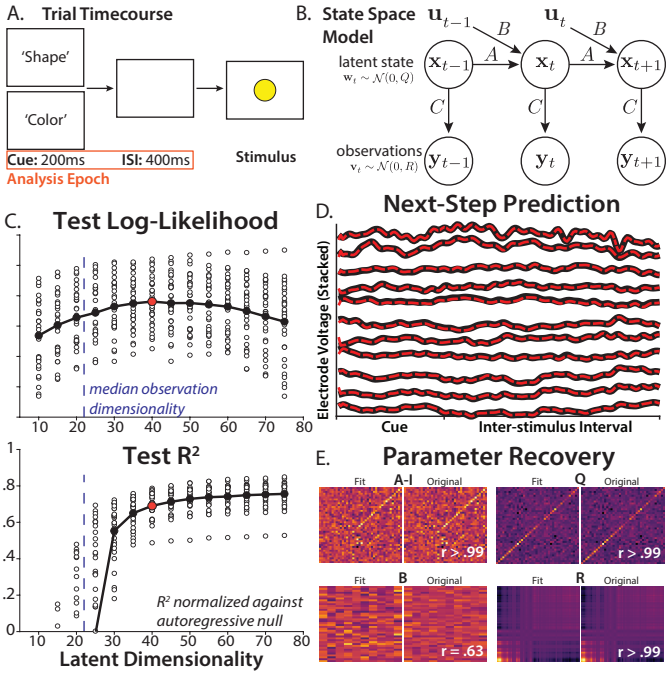


Figure 1: *Model validation.* **A)** On each trial, participants performed a pre-cued task (50% switch rate). **B)** We modeled the latent timeseries $\mathbf{x}_{1:T}$ that gives rise to our electrode timeseries $\mathbf{y}_{1:T}$, assuming linear dynamics, linear observations, and Gaussian noise. **C)** *Top:* The best-fitting models had more latent dimensions than observed dimensions. *Bottom:* SSMs fit better than vector autoregressive (VAR) models fit directly to the observations. **D)** Next-timestep prediction for a single held-out trial. Black line: EEG voltage, red dashed line: model prediction. **E)** Accurate recovery of SSM parameters (only a subset shown, but good performance throughout).

**SSM validation.** We chose the number of latent dimensions ($\dim_{\mathbf{x}}$) through cross-validation. The best-fitting $\dim_{\mathbf{x}}$was substantially larger than the effective observation dimensionality ($\dim_{\mathbf{y}}$; Fig 1C). Fitted SSMs could accurately filter held-out data (Fig 1D), with much better accuracy than autoregressive models fit directly to the component timeseries. Despite high $\dim_{\mathbf{x}}$, we could accurately recover parameters from simulated data (up to invertible transformation; Fig 1E).

**Neural indices of task control** Plotting the posterior estimates of latent state, we found that neural states for each task separated over time after the cue (Fig 2A).

We wanted to quantify how task-related neural dynamics change under the hypothesized deployment of cognitive control during task switching. To answer this question, we turned to a classical control theoretic metric of *controllability* (Kao & Hennequin, 2019). We developed a modified form of the con-

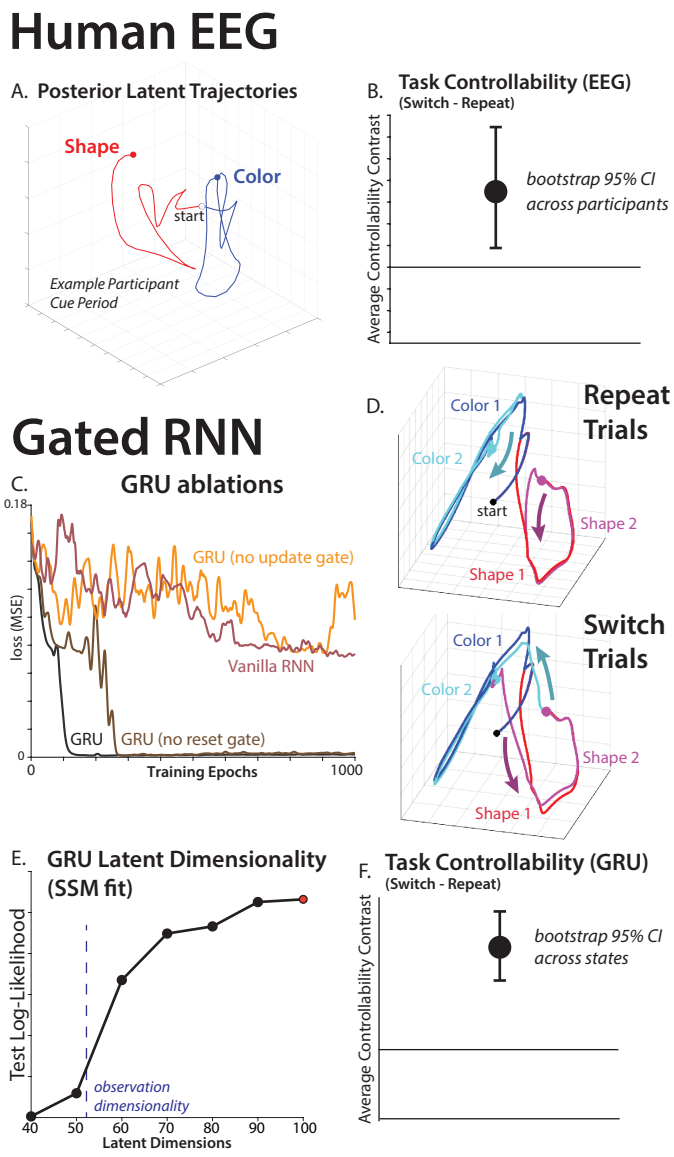


Figure 2: *Neural signatures of task control.* **A)** Posterior latent state estimates, averaged over trials in each task (red vs blue) for an example participants. Projected onto the top eigenvectors of $\mathcal{W}_c$ (i.e. the controllable subspace) and smoothed for visualisation. **B)** Average Task $\mathcal{W}_c$ was significantly higher for switch and repeat trials. **C)** GRU fit across different variants. Fit is assessed on noiseless inputs, whereas the model is fit to noisy inputs. Ablating the reset gate (fixing to open) has little cost to performance. In contrast, ablating the update gate (fixing to open) prevents the network from mastering the task, resulting in similarly poor performance as un-gated RNNs. **D)** PCA projection of hidden unit activations, averaged over trials. Red/blue lines indicates the first trial (Shape 1 & Color 1), magenta/cyan lines indicates the second trials (Shape 2 & Color 2). Note that in repeat trials (top) trial 2 stays in the same location at trial 1, whereas in switch trials (bottom) they swap locations. **E)** Similar to participants, SSMs fit to GRU hidden unit activations showed better test liklihoods at high latent dimensionality. **F)** Similar to participants, average Task $\mathcal{W}_c$ in GRUs was significantly higher for switch trials than repeat trials.

trollability gramian ($\mathcal{W}_c$), which indexes the asymptotic spread of input energy throughout a system:

$$\mathcal{W}_c = (\sum_{t=0}^{\infty} A^t Q A^{t\top})^{-1} \sum_{t=0}^{\infty} A^t B B^\top A^{t\top}$$

We whitened the gramian with the estimated asymptotic state noise (inverse term). This normalization accounts for noise, and makes our metric invariant across the SSM solution manifold (see also: Bouchard & Kumar, 2024). We computed $\mathcal{W}_c$ by solving the corresponding discrete Lyapunov functions. To isolate the effect of *task* in particular, we only used the columns of $B$ that coded task identity ($\mathbf{b}_{\text{task}}$). To test how task-based control met the cognitive demands of task switching, we contrasted the 'average task controllability', $\texttt{trace}(\mathcal{W}_c)$, between switch and repeat trials.

We found that 'task control' was stronger on switch trials than repeat trials (permutation $p = .019$; Fig 2E), consistent with the active reconfiguration of neural states to implement a new task. While incentives and subsequent RTs were encoded during this period (not shown), we did not find strong evidence that these factors modulated task control.

**Gated recurrent unit (GRU) model.** To understand whether optimized systems show our putative indices of task control, we next explored how gated recurrent neural networks switch between similar tasks (Cho et al., 2014). Using PyTorch, we trained a GRU ($1 \times 108$ hidden units) on epochs containing sequential pairs of trials. Mimicking the behavioral experiment, the network received (noisy) inputs for the task cue and the stimulus features.

We found that GRUs quickly learn this task (Fig 2C). Gate ablations revealed a key role for the 'update' gate that controls integration timescales (Krishnamurthy, Can, & Schwab, 2022), a potential mechanism for changing the previous task state.

**Task control in GRUs.** We visualized hidden unit activations using PCA, finding that GRUs learned distinct representations for each task, regardless of trial order (Fig 2D).

To quantify these dynamics, we fit our SSM to the GRU hidden unit activations during the second cue period using subspace identification. We included inputs for *task*, *switch*, and *task*$\times$*switch*. Mirroring participants, the best $\dim_{\mathbf{x}}$ was greater than the PCA-reduced $\dim_{\mathbf{y}}$ (Fig 2E) and our linear model could accurately predict GRU dynamics ($R^2_{CV} = .90$).

In our critical test, we compared the GRU's average task $\mathcal{W}_c$ between switch and repeat trials. The GRU showed the same key signature as participants: higher average task $\mathcal{W}_c$ when switching tasks than when repeating tasks (Fig 2F).

**Conclusions.** Our analyses reveal a control theoretic signature of how natural and artificial neural systems dynamically reconfigure task processing. Speculatively, this index of neural control may reflect the deployment of attention to gate task-relevant representations (Braver & Cohen, 1999). Future work could extend these methods to explore source-localized neural dynamics, and to explicitly test theories of optimal feedback control over neural states (Ritz, Leng, & Shenhav, 2022).

## References


Bouchard, K., & Kumar, A. (2024, March). Feedback controllability is a normative theory of neural population dynamics. *Research Square*.

Braver, T. S., & Cohen, J. D. (1999). Dopamine, cognitive control, and schizophrenia: the gating model. *Prog. Brain Res.*, *121*, 327–349.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, June). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv [cs.CL]*.

Hall-McMaster, S., Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2019, October). Reward boosts neural coding of task rules to optimize cognitive flexibility. *J. Neurosci.*, *39*(43), 8549–8561.

Jaffe, P. I., Poldrack, R. A., Schafer, R. J., & Bissett, P. G. (2023, January). Modelling human behaviour in cognitive tasks with latent dynamical systems. *Nature Human Behaviour*, 1–15.

Kao, T.-C., & Hennequin, G. (2019, September). Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics. *Curr. Opin. Neurobiol.*, *58*, 122–129.

Krishnamurthy, K., Can, T., & Schwab, D. J. (2022, January). Theory of gating in recurrent neural networks. *Phys. Rev. X*, *12*(1), 011011.

Linderman, S., Nichols, A., Blei, D., Zimmer, M., & Paninski, L. (2019). Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in c. elegans. *BioRxiv*.

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

Musslick, S., Jang, S. J., Shvartsman, M., Shenhav, A., & Cohen, J. D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. In *CogSci.* shenhavlab.org.

Ritz, H., Leng, X., & Shenhav, A. (2022, March). Cognitive control as a multivariate optimization problem. *J. Cogn. Neurosci.*, *34*(4), 569–591.

Smith, A. C., & Brown, E. N. (2003, May). Estimating a state-space model from point process observations. *Neural Comput.*, *15*(5), 965–991.

Stone, I. R., Sagiv, Y., Park, I. M., & Pillow, J. W. (2023, April). Spectral learning of bernoulli linear dynamical systems models for decision-making. *Transactions on Machine Learning Research*.