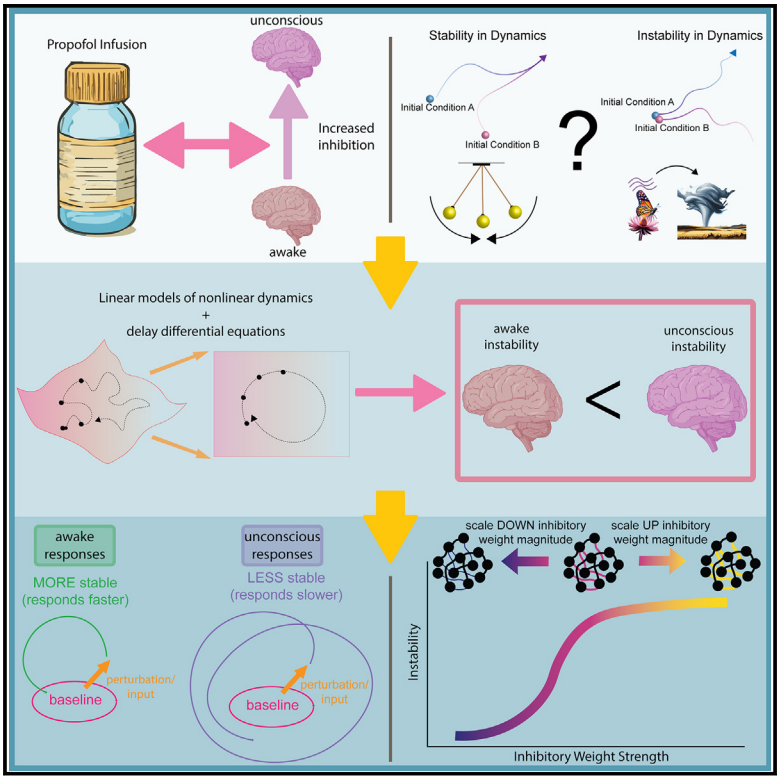
Neuron

Article

Propofol anesthesia destabilizes neural dynamics across cortex

Graphical abstract



Highlights

- We developed DeLASE, a method for quantifying changes in neural stability
- During propofol-induced unconsciousness, neural activity was destabilized
- Destabilized artificial systems had similar dynamics to the destabilized brain
- Increasing inhibition, as propofol does, destabilized artificial network activity

Authors

Adam J. Eisen, Leo Kozachkov, André M. Bastos, ..., Emery N. Brown, Ila R. Fiete, Earl K. Miller

Correspondence

fiete@mit.edu (I.R.F.), ekmiller@mit.edu (E.K.M.)

In brief

Eisen and Kozachkov et al. develop a method to measure changes in neural stability. They find evidence that an anesthetic causes unconsciousness by destabilizing neural activity. It makes activity more susceptible to perturbation. This is due to excess inhibition in the brain.

Eisen et al., 2024, Neuron *112*, 1–15
August 21, 2024
https://doi.org/10.1016/j.neuron.2024.06.011

CellPress

Neuron

CellPress
OPEN ACCESS

Article

Propofol anesthesia destabilizes neural dynamics across cortex

Adam J. Eisen,[1,2,3,4,11] Leo Kozachkov,[2,3,4,11] André M. Bastos,[5,6] Jacob A. Donoghue,[3,7] Meredith K. Mahnke,[1] Scott L. Brincat,[1,3] Sarthak Chandra,[2,3,4] John Tauber,[8] Emery N. Brown,[1,3,9,10] Ila R. Fiete,[2,3,4,12,*] and Earl K. Miller[1,3,12,13,*]

[1]The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]The K. Lisa Yang Integrative Computational Neuroscience Center, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Department of Psychology, Vanderbilt University, Nashville, TN 37235, USA
[6]Vanderbilt Brain Institute, Vanderbilt University, Nashville, TN 37235, USA
[7]Beacon Biosignals, Boston, MA 02114, USA
[8]Department of Mathematics and Statistics, Boston University, Boston, MA 02215, USA
[9]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[10]Division of Sleep Medicine, Harvard Medical School, Boston, MA 02115, USA
[11]These authors contributed equally
[12]Senior author
[13]Lead contact
*Correspondence: fiete@mit.edu (I.R.F.), ekmiller@mit.edu (E.K.M.)
https://doi.org/10.1016/j.neuron.2024.06.011

SUMMARY

Every day, hundreds of thousands of people undergo general anesthesia. One hypothesis is that anesthesia disrupts dynamic stability—the ability of the brain to balance excitability with the need to be stable and controllable. To test this hypothesis, we developed a method for quantifying changes in population-level dynamic stability in complex systems: delayed linear analysis for stability estimation (DeLASE). Propofol was used to transition animals between the awake state and anesthetized unconsciousness. DeLASE was applied to macaque cortex local field potentials (LFPs). We found that neural dynamics were more unstable in unconsciousness compared with the awake state. Cortical trajectories mirrored predictions from destabilized linear systems. We mimicked the effect of propofol in simulated neural networks by increasing inhibitory tone. This in turn destabilized the networks, as observed in the neural data. Our results suggest that anesthesia disrupts dynamical stability that is required for consciousness.

INTRODUCTION

The pharmacological action and neurophysiological response of propofol are well understood, but how it renders unconsciousness is not. Propofol boosts inhibition through $GABA_A$ (γ-aminobutyric acid type A) receptors and significantly alters cortical dynamics.[1–6] This could disrupt the cortical communication on which consciousness depends,[1,2,7] but the exact link to theories of consciousness is not clear. Many theories of consciousness have focused on the representation and network structure involved in integrating information or linking together cortical representations.[8–14] For example, one prominent theory of consciousness posits that awareness follows from an "ignition" that produces widespread cortical spiking, much like a few claps can lead to a whole audience applauding.[10,15,16] However, overly excitable and unstable states are uncontrollable, indicative of pathological conditions.[17,18] Thus, we hypothesize that a key factor in consciousness is *dynamic stability*. Brain states should be sufficiently excitable for generation of widespread activity and information integration. But they also need to be controllable and stable, reliably producing the same computations.[19–23]

Stability has long been known to be critical for brain function, but early computational work investigated it in the context of convergence to a single state, involving a kind of "freezing" of neural activity.[24–26] However, normal neural activity is rarely so stationary; rather, it constantly evolves through dynamic trajectories.[27,28] Thus, stability, and hence consciousness, needs to be understood in terms of a dynamic brain.[21,22] Here, we approach the analysis of anesthetic unconsciousness through the lens of *dynamic stability* (henceforth stability) (Figure 1A), a fundamental concept in dynamical systems theory and control. Essentially, dynamic stability is a measure of the robustness of a dynamical system. The system needs to be able to recover from disturbances (e.g., distractions, random fluctuations in activity) to its normal state.

Previous work on cortical stability during anesthesia has produced contradictory results, suggesting that anesthesia either

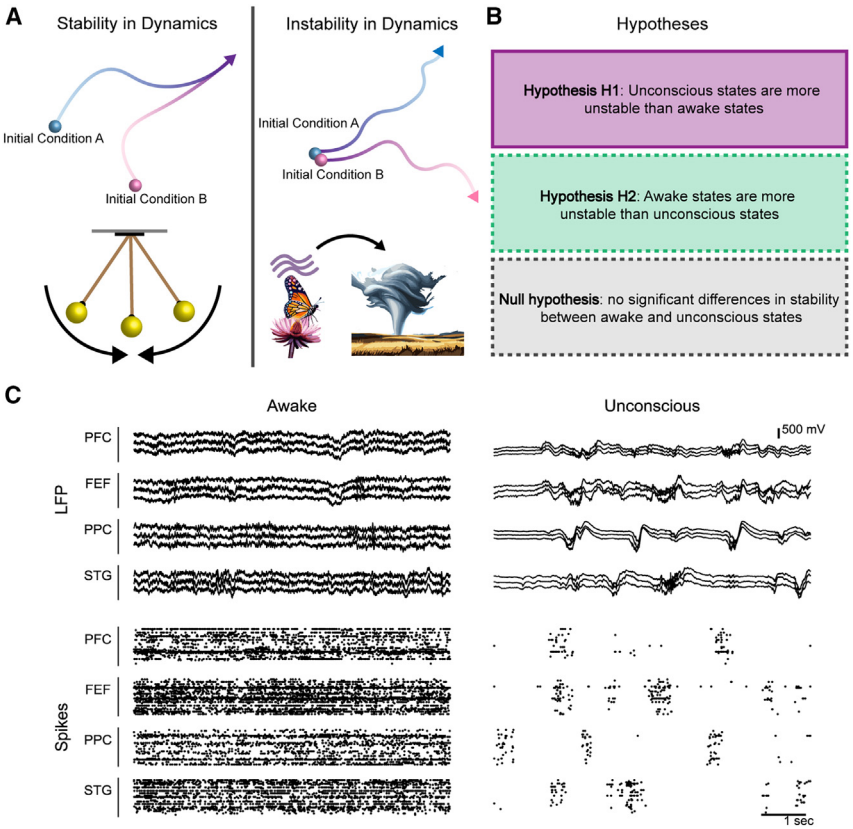


Figure 1. Introduction: Stability and instability, and hypothesis candidates

(A) (Left, top) Depiction of stability in dynamics: starting from two distinct initial conditions, system trajectories converge. (Left, bottom) Diagram of a stable system—a pendulum with friction—that will converge to the bottom position regardless of the starting point. (Right, top) Depiction of instability in dynamics: starting from two similar initial conditions, system trajectories diverge to distinct paths. (Right, bottom) Cartoon of an unstable system—the weather—where a small perturbation, like a butterfly's wings flapping, may cause a large-scale change such as a tornado.[29]

(B) Three hypotheses regarding the impact of propofol anesthesia on neural dynamics: dynamics can be more unstable, more stable, or show no significant change compared with awake dynamics.

(C) Sample neural data from the propofol dataset: (top row) LFPs during the awake (left) and unconscious (right) states. (Bottom row) Spike rasters during awake (left) and unconscious (right) states. In the awake state, LFP signals are lower amplitude with higher-frequency activity, and spiking is consistent without coordinated bursting. In the unconscious state, LFP signals display low frequency, hypothesized to underlie loss of consciousness.[1,2] Spiking during the unconscious state shows up-state/down-state bursting patterns.

destabilizes[30,31] or excessively stabilizes[32–34] neural dynamics (Figure 1B). This could be due to a paucity of studies using high-density intracortical electrophysiology and the inability to therefore apply sufficiently rich dynamical tools to assess stability. Thus, we used a dataset of local field potential (LFP) recordings with hundreds of electrodes from multiple brain regions in two non-human primates (NHPs, specifically adult rhesus macaque monkeys) as they lost and regained consciousness due to propofol anesthesia (Figure 1C).

We introduce a new approach—delayed linear analysis for stability estimation (DeLASE). DeLASE directly quantifies changes in stability in neural data. We show that this method produces high-quality models of nonlinear circuit dynamics while maintaining the simplicity and tractability of a linear dynamical system. We validated the model's estimates of changes in dynamic stability in systems for which the ground truth stability is known. We found that propofol-induced unconsciousness is associated with destabilized neural dynamics.

RESULTS

Dynamical systems approach: DeLASE method

There are three major challenges to a dynamical systems analysis of neural data.

(1) Partial observation: Neural dynamics are high-dimensional. Any one sample of neural data can only hope to capture a few of the dimensions of the overall system. This means we are restricted to partial observation.
(2) High-dimensionality: These partial observations are nonetheless high-dimensional. Electrophysiology can employ hundreds of electrodes and a high LFP sampling rate. This generates thousands of neural data points every second. Therefore, analysis of neural dynamics requires a computationally efficient model-fitting procedure to handle vast amounts of data.
(3) Nonlinearity: Neural dynamics are highly nonlinear. Nonlinear models are computationally intensive. Any model must be optimized for computational tractability while still capturing nonlinear dynamics.

In this section, we propose a new method that leverages advances in data-driven dynamical systems analysis with strong theoretical guarantees to solve each of these challenges.

Partial observations

Takens' Delay Embedding Theorem is a powerful result from dynamical systems theory that shows how the full attractor of a dynamical system can be reconstructed from partial observation of data by appending lagged copies of the subset of observed variables onto the existing measurements.[35] In other words, remarkably, it is possible to trade time (multiple time observations of a subset of variables) for space (the full attractor of the system) to characterize a nonlinear dynamical system. As an example of this delay embedding principle,[36–47] the famous

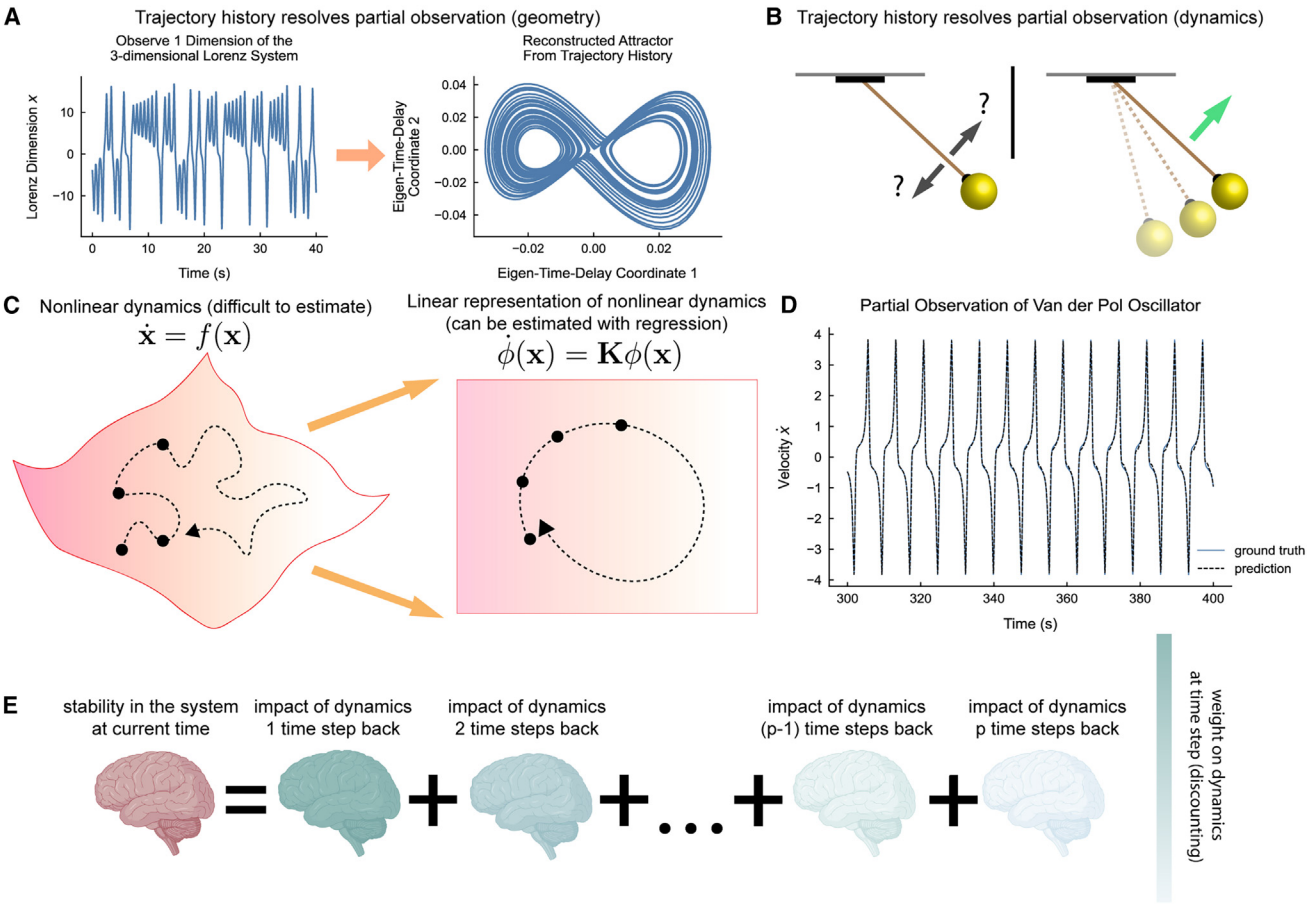


Figure 2. DeLASE: Measuring stability through linear delayed dynamical models

(A) Trajectory history enables attractor reconstruction from partial observation. (Left) With only a single observation of the three-dimensional Lorenz system, it is not clear what the attractor geometry is. (Right) However, by performing a delay embedding, the famous butterfly attractor geometry can be reconstructed. We plot the first two eigen-time-delay coordinates of the data, equivalent to performing principal-component analysis (PCA) whitening on the delay embedding matrix.

(B) Trajectory history enables prediction in partially observed systems. (Left) With only the position of a pendulum, no prediction of future states can be made. (Right) The trajectory history illuminates the impact of the unobserved variable (the pendulum velocity) on the observed variable (the position).

(C) A cartoon depiction of the Koopman operator. A flow on a nonlinear manifold can be expanded (in this case through time-delay embedding) into a high-dimensional embedding in which there exists a linear representation for the dynamics.

(D) Dynamical modeling of the partially observed second dimension of the nonlinear Van der Pol oscillator. The HAVOK model was able to achieve a fully autonomous linear representation of the nonlinear oscillator.

(E) A cartoon of the approach for assessing stability from delay differential equations is illustrated. The impacts of modeled dynamics are weighted based on how far back they are from the current state.

"butterfly attractor"[29] of the three-dimensional Lorenz system can be reconstructed from observing a single dimension over the full time period (Figure 2A). The key takeaway is that information is gained through considering the trajectory history of a system, in conjunction with its current state. As we hope to convey in this paper, this has tremendous implications for neural data, which is likely capturing a very small fraction of the information contained in the full neural system. Delay embeddings have been widely harnessed as a tool for elucidating dynamical insights from partially observed data,[48–55] including examples in the context of neural data.[42,56–61]

The trajectory history also enables reconstruction of the time-resolved dynamics from partial observations.[57,62–67] For instance, if we observe a snapshot of a pendulum's position but not its velocity (a partial observation), we cannot predict its next state (Figure 2B). By including the position history, the upcoming dynamics of the pendulum become clear. We harness the information provided by the trajectory history of neural data in a similar way to predict future states.

High-dimensionality and nonlinearity

There are many efficient tools for accurate prediction of linear systems. However, neural and neural circuit dynamics are highly nonlinear, and nonlinear prediction is often both challenging and computationally expensive. Here, we leverage a second deep insight from dynamical systems, the Koopman operator theory, which shows that a *nonlinear* dynamical system can be represented without approximation as an *infinite-dimensional linear* system[68–70] (Figure 2C). The major challenge in practically

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

exploiting this theoretical insight is to find large but *finite*-dimensional representations that allow the Koopman operator theory to approximately hold.

Given that incorporating trajectory history through delay embedding reconstructs the underlying attractor from partial observation, we might surmise that the dimension-expansion from delay embeddings constitutes a reasonable finite-dimensional representation in which the Koopman theory approximately holds.[36,37,41,42,71–73] An approach that exploits exactly this insight is the Hankel alternative view of Koopman (HAVOK),[36] which uses a decorrelated and low-rank representation of the delay embedding matrix (known as eigen-time-delay coordinates) as an embedding space in which to estimate the Koopman operator. The resulting decorrelated representation used to estimate the dynamics is relevant for neural data, which can often be highly correlated. Under certain conditions, the HAVOK approach has been shown to find an optimal finite-dimensional space for representing the Koopman operator.[37] HAVOK is a variant of dynamic mode decomposition (DMD),[74–76] an approach to estimating the Koopman operator that has been explored in many varieties,[71,74,77–86] including applications to neuroscience.[42,86–90] To demonstrate the predictive power of HAVOK models, we take a partial observation of the Van der Pol oscillator (a two-dimensional nonlinear system).[91] HAVOK models are able to autonomously reproduce this nonlinear time series with purely linear dynamics (Figure 2D). By "autonomously reproduce," we mean that future predictions are generated by previous ones.

Therefore, we use HAVOK to construct efficient and accurate dynamical models of partially observed neural circuits, resolving the challenges of high-dimensionality and nonlinearity.

Estimating stability from delayed dynamical systems models

We use the accurate dynamical models constructed with HAVOK to estimate the stability of the observed dynamics. We rearrange the dynamics equation, shifting from a Koopman representation in eigen-time-delay coordinates to a delay differential equation in the original neural space. Delay differential equations make explicit the dependence of the future states of the systems on past states. The stability of a system described by a delay differential equation is determined by the roots of its corresponding characteristic equation.[92]

These roots, known as characteristic roots, are complex-valued numbers. The real part corresponds to the (inverse) characteristic timescale at which the system will respond to a perturbation along a particular direction. The sign in front of the timescale determines whether the system will grow (a positive sign) or decay (a negative sign) in response to a perturbation. For instance, a negative root of $-100\ s^{-1}$ will decay quickly in response to perturbation (in this case, with a characteristic timescale of 0.01 s). A negative root of $-10\ s^{-1}$ is more unstable than a negative root of $-100\ s^{-1}$ because the response will decay more slowly—namely, with a timescale of 0.1 s. The imaginary part of the characteristic root is the frequency of the perturbation response.

To numerically approximate a finite portion of the (infinitely many) roots of a given delay differential equation, we harness the TRACE-DDE (tool for robust analysis and characteristic equations for delay differential equations) algorithm.[93] This algorithm broadly estimates stability by discounting the impact of the previous time steps based on how far back they are from the current time (Figure 2E).

DeLASE, our approach to directly estimating changes in stability in neural data, consists of four primary steps:

(1) Performing a grid search across key HAVOK hyperparameters (the size of the delay embedding, the rank of the eigen-time-delay coordinates) to identify the optimal parameters for all the dynamic states being compared.
(2) Fitting HAVOK dynamical systems models to the data.
(3) Using the TRACE-DDE algorithm to extract characteristic roots from the delay differential equations representation of the models as an estimate of stability.
(4) Compare the estimated stability from each of the dynamic states.

DeLASE tracks changes in stability in simulated neural networks from partial observations

To validate the DeLASE procedure, we attempted to accurately predict changes in stability from partial observations of systems for which the ground truth stability is known.

We began by considering noise-driven linear dynamical systems of varying stability. We picked 10 different degrees of stability of such systems, all of which were close to the transition to unstable dynamics (as awake and unconscious neural states are hypothesized to be).[19,30,32,94] We altered the degree of stability by varying the maximum eigenvalue of the linear system matrix. For the stability analysis, we randomly chose 10% of the dimensions of the system for partial observation and fit HAVOK models. The correlation of the mean instability estimated from DeLASE with the ground truth stability is 1 (Figure 3A).

To further validate the DeLASE procedure on more brain-like systems, we considered numerous randomly connected recurrent neural networks (RNNs) with a gain parameter g. The gain parameter scales the synaptic weights and induces a transition to chaos in these networks.[97] We ran DeLASE on a randomly chosen partial observation of approximately 1% of the dimensions from each network. DeLASE predicted relative changes in the stability of RNNs with a correlation of 0.993 (Figure 3B). For reference, we compare DeLASE to vector autoregression (VAR), a simple linear dynamical systems model, which has been used previously to estimate stability in neural data.[32,33,98] Both methods perform somewhat comparably on stability estimation where ground truth is known (i.e., on simulated data). However, on actual neural data, DeLASE consistently outperforms VAR on next time-step prediction. This is explained in the following section.

Delayed linear dynamical systems models capture neural dynamics

Good estimates of stability depend on good models of dynamics. In other words, if we are to trust a statistical method for estimating dynamical variables from data via model fitting, then that model should do a good job of predicting the future data. We chose HAVOK models because they use a history of

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

CellPress
OPEN ACCESS

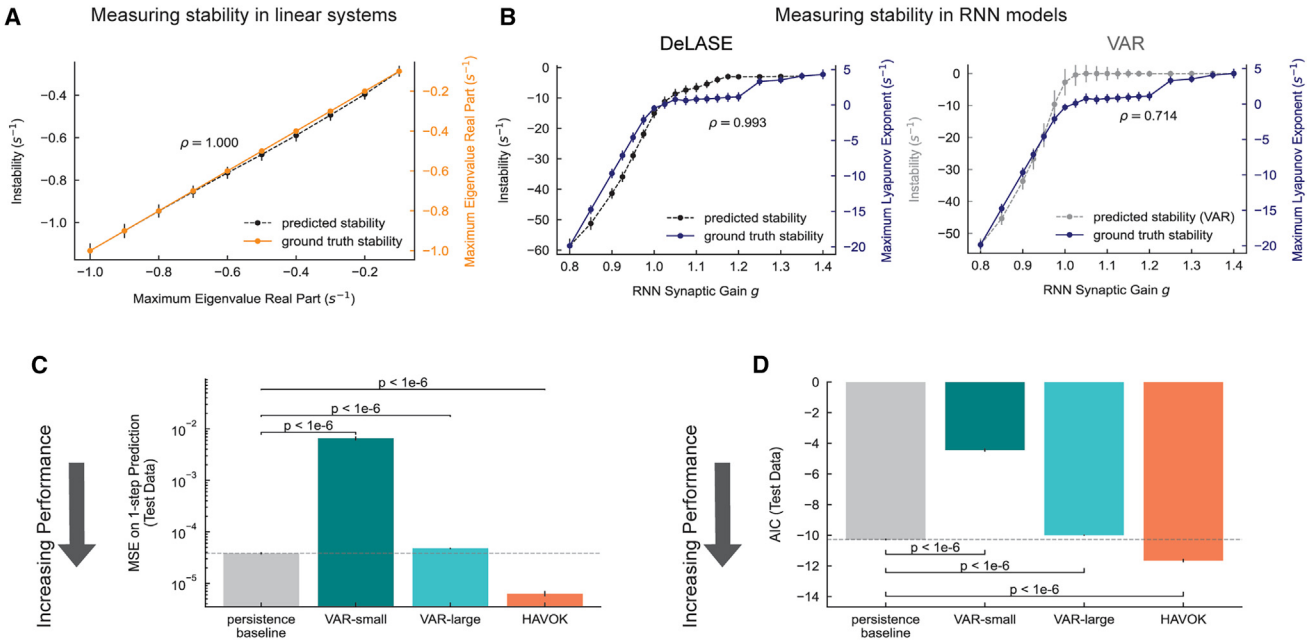


Figure 3. Validating DeLASE on simulated and empirical neural data

Data are represented as mean ± SEM.

(A) Linear systems. The x axis and right y axis show the maximal real part of the dynamics matrix eigenvalues (orange line). The left y axis displays instability estimated from DeLASE (black dotted line), using the maximal 10% of the real parts of characteristic roots from delay differential equations analysis (averaged over 10 simulations). DeLASE was fit to 10 out of 100 system dimensions and achieved a Pearson correlation of 1.

(B) RNN models. (Left) The x axis is the gain parameter, which scales synaptic weights and increases network instability. The right y axis shows the maximum Lyapunov exponent (blue curve), averaged over 10 simulations, determining stability. The full system state and dynamics are needed to compute this in the standard approach we used.[95] The left y axis displays instability estimated from DeLASE (black dotted line), with the same approach as (A). DeLASE was fit 10 out of 1,024 system dimensions and achieved a Pearson correlation of 0.993. (Right) The VAR stability estimation approach is shown, fit to a large temporal window, and using the maximal 10% of real parts of eigenvalues. VAR was fit to 10 out of 1,024 dimensions and achieved a Pearson correlation of 0.714. Both methods struggle between synaptic gains of 1 and 1.2, where networks transition from quasiperiodic to chaotic dynamics, which are difficult to distinguish.[96]

(C) Comparison of different models' one-step predictions on neural data (LFPs from four recorded areas, sampled at 1 kHz). The persistence baseline, predicting the next state as identical to the previous one, is plotted as the gray bar and dotted line. HAVOK models were the only ones to surpass this baseline in prediction tasks.

(D) Similar to (C), but using AIC, showing that HAVOK's superior performance was not due to having more parameters than other models.

See also Figure S1.

neural states to construct simpler linear dynamics from a nonlinear system (like the brain). Our first step toward estimating changes in stability in real neural data, therefore, was to confirm that HAVOK was a good model for capturing neural dynamics.

We compared HAVOK to three other models. (1) A persistence baseline model. This model predicts that the neural state at each time step will be identical to the previous time step. For instance, given a time series with values: [1, 4, 5, 7], when predicting the second time point, the persistence baseline is 1, and when predicting the third one, the persistence baseline is 4. Better models of neural dynamics should thus be able to outperform the persistence baseline model. (2) and (3) Two forms of VAR models, which were previously used to study stability in propofol-induced unconsciousness.[32,33] Here, we use 1st-order VAR (VAR(1)) models, which generate predictions for the next state using only the most recent state but also take into account the dynamics across the training set. One VAR model (VAR-small) used 500 ms of neural data for training, as in previous work.[32] The other VAR model (VAR-large) used a 15 s window. VAR-large was included because we used 15 s windows for HAVOK. All models were tested on a window of data of equal size to the training window and temporally immediately following the training window. HAVOK, in contrast to VAR, predicts future states by taking into account the history of multiple preceding timesteps.

We computed the mean-squared error (MSE) of one-step model predictions averaged over all sessions (Figure 3C). The VAR-small models were significantly outperformed by the persistence baseline models ($p < 1e{-}6$, one-sided Wilcoxon signed-rank test). The VAR-large models achieved an MSE much closer to the persistence baseline but were still outperformed ($p < 1e{-}6$, one-sided Wilcoxon signed-rank test). Only the HAVOK models were capable of beating the persistence baseline ($p < 1e{-}6$, one-sided Wilcoxon signed-rank test). Akaike information criterion (AIC) was also used to assess the models' prediction quality relative to the number of parameters. AIC penalizes models for complexity and thus guards against overfitting data. HAVOK models have more parameters. Again, only the HAVOK models outperformed the persistence baseline models

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

on one-step prediction (Figure 3D, $p < 1e{-}6$, one-sided Wilcoxon signed-rank test). All results held when considering the prediction quality for each NHP separately (Figure S1).

Thus, through the inclusion of the history of neural states, HAVOK models form accurate dynamical models of the multi-electrode activity while preserving the tractability and interpretability of linear methods.

Propofol anesthesia destabilizes cortical neural dynamics

To determine the impact of propofol anesthesia on the stability of neural dynamics, we analyzed multi-electrode activity recorded from two NHPs.[2] Electrodes were placed in four areas: ventrolateral prefrontal cortex, frontal eye fields, posterior parietal cortex, and auditory cortex. We found that propofol destabilized neural activity.

We first characterized the stability of every brain region separately using DeLASE (Figure 4A). Characteristic roots were used as a quantification of instability. One component is the timescale at which the system will respond to a perturbation along a particular direction (see "estimating stability from delayed dynamical systems models"). The faster the response, the more stable the system. While we later analyze neural responses to experimental perturbations in the form of stimuli, these perturbations are hypothetical and used to describe the general system response. We analyzed the instability values from the upper 10% of the distribution of characteristic roots (Figure 4A). We used the upper 10% because they have the most impact on dynamic stability. For both animals and across all four cortical regions, propofol anesthesia reliably destabilized neural dynamics (blue and gray curves, $p < 0.001$ for all combinations of NHP/area, one-sided Wilcoxon signed-rank test). The same was true for all areas considered together as a single system (purple curves, $p < 0.001$ for both NHPs, one-sided Wilcoxon signed-rank test). Note that for each curve, the instability values increased after propofol induction. This means the system was slower to respond (i.e., less stable).

Note that the absolute values of instability were variable across areas and NHPs. Because of the nature of the instability values, their absolute value will capture details of the differences in signals due to factors such as differences in their intrinsic dynamics, exact location of the electrodes relative to the neurons, electrode impedance, etc. The critical measure, therefore, is the change in values over time. When normalized to baseline values, the change in instability across areas differed in magnitude but followed similar time courses (Figure S2A). When we considered all recorded areas as a single system, the ratio of change in instability values was remarkably similar across the two NHPs (Figure 4B). For the remainder of this paper, we focus on models constructed from all areas as a single system (Figure 4B). Instability values during unconsciousness were nearly twice that of awake states (pre-propofol), indicating that the system's responses to perturbation were much slower than the response during the awake state (Figure 4B).

To better visualize the change in instability values, we plot the distribution of the top 10% of values at each time point across all areas (Figure 4C). The entire distribution of instability values shifted upward during propofol infusion ($p < 0.01$ for both NHPs, Wilcoxon signed-rank test on the Cramér-von Mises criterion of awake-unconscious distributions for each session). This was also true when considering the instability value distribution below the top 10% (Figure S2B). During recovery, the distribution of values shifted down to be more stable ($p < 0.001$, one-sided Wilcoxon signed-rank test), approaching that seen pre-propofol (Figure S2C, $p < 0.001$ for both NHPs, one-sided Wilcoxon signed-rank test on the Cramér-von Mises criterion of awake-recovery distributions compared with unconscious-recovery for each session).

Instability changed throughout the session. It rose to a maximum during the large loading dose, then gradually declined through the maintenance dose and finally recovery phases (Figure 4D). The estimated integrated dosage of propofol was predictive of instability in the system (Figure S2D, R^2 of 0.873 for NHP 1 and 0.853 for NHP 2 using a linear-log model).

Next, we considered another measure of system response to perturbation, the frequency of the system response (Figure 4E). During unconsciousness, a larger portion of the frequency components fell in the lower-frequency bands than during the awake state (Figure S2E, $p < 0.01$ for both NHPs, Wilcoxon signed-rank test on the proportion of frequencies in the delta band). Conversely, during the awake state, more fell into the highest populated frequency bands (Figure S2E, $p < 0.01$ for both NHPs, Wilcoxon signed-rank test on the proportion of frequencies in the beta band for NHP 1 and gamma band and NHP 2). This is consistent with observations that propofol causes a large increase in lower-frequency power and decrease in higher-frequency power in cortex.[2]

Destabilized dynamics explain sensory responses in the anesthetized cortex

In the previous section, we established that propofol measurably destabilizes cortical neural dynamics. We now investigate how this destabilization changes sensory responses. We found that sensory responses under anesthesia are consistent with destabilized linear filters.

When more stable systems are perturbed, they recover faster than less stable systems. An example of this is two systems in which spheres of different mass are hanging at rest from identical springs (Figure 5A). A smaller mass causes the system to be more stable. Thus, when perturbed with identical force, the spring with the smaller mass decays back to its resting state more quickly.

We can describe the pattern of response to perturbation with a simple linear filter of the form $\dot{x} = -\lambda x + u(t)$, where $u(t)$ is a perturbation. In this filter, the parameter λ controls the "stability" of the filter. For larger values of λ, the filter is more stable in that the filter will take less time to recover in response to perturbations. We considered the response of such a filter to a pure sinusoidal input ($u(t) = sin(t)$) across a wide range of λ values (Figure 5B). Smaller values of λ corresponded to larger amplitude, phase-shifted responses. To illustrate such responses in a more neural-like setting, we simulated two different linear filters, driven by a small amount of noise (Figure 5C). At times 0 and 1, sinusoidal inputs were provided to the linear filter. The purple curve—having a less stable parameter λ—shows a response with a phase shift relative to the green curve, which has a more stable parameter.

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

CellPress
OPEN ACCESS

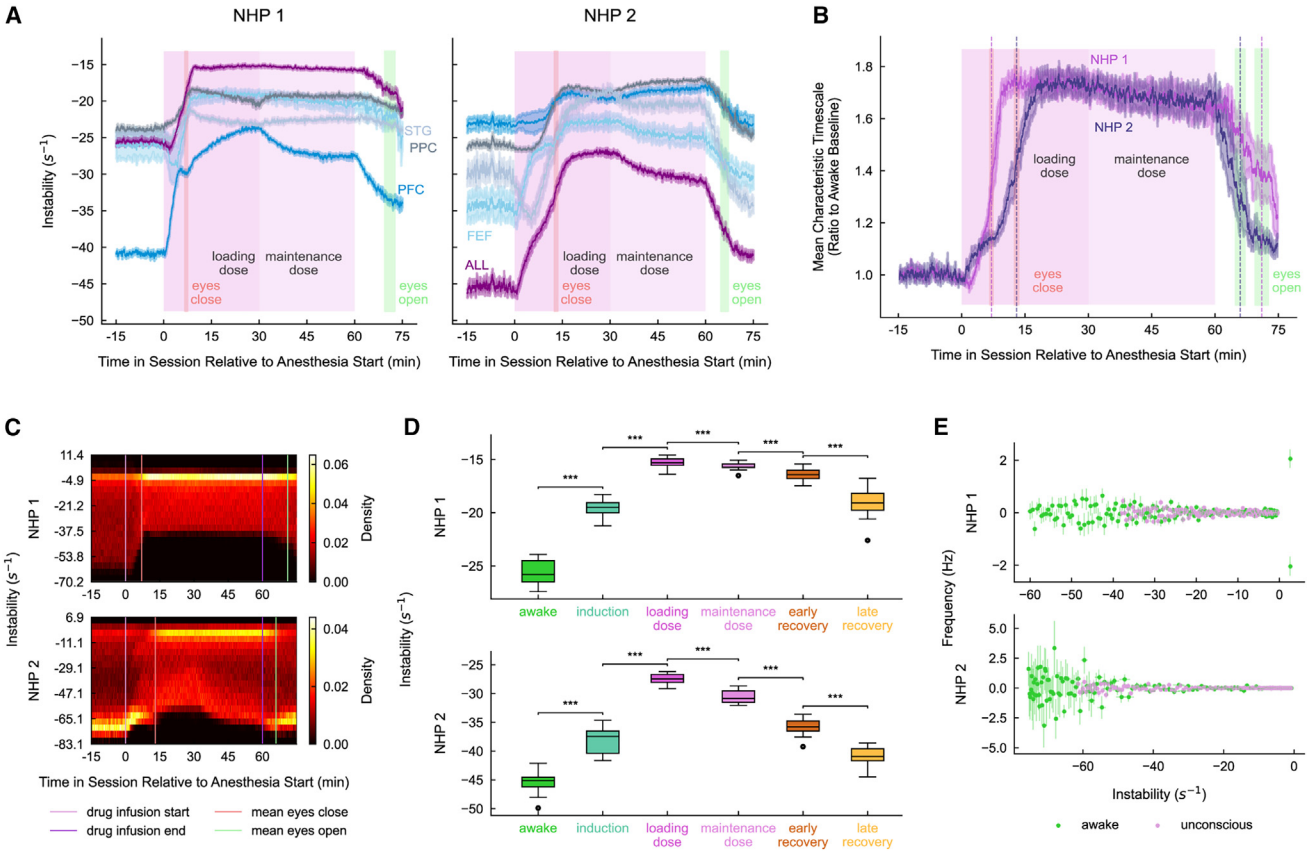


Figure 4. Propofol anesthesia destabilizes cortical neural dynamics

All characteristic root analyses refer to the maximal 10% of the distribution of characteristic roots extracted from delay differential equations analysis. All figures (except C) are reported as mean ± SEM, with results averaged over sessions for each NHP.

(A) Instability across the session in both NHPs. DeLASE was fit to contiguous windows of 15 s across the session for each area individually. The x axis is time relative to the start of anesthesia. The y axis is the mean real part of characteristic roots (the inverse timescale of response)—a measure of instability. Instability increased during anesthesia for all areas in both NHPs. The red line indicates the mean moment of eyes closing (used as a proxy for loss of consciousness), and the green line indicates the mean moment of eyes opening (used as a proxy for return of consciousness). Areas are ventrolateral prefrontal cortex (PFC), frontal eye fields (FEFs), posterior parietal cortex (PPC), and auditory cortex (STG). "ALL" refers to all areas considered together as a single system.

(B) The mean timescales of response normalized to the awake baseline for "ALL" areas considered together. The awake baseline was computed for each session by taking the geometric mean of the timescales associated with the characteristic roots across all windows in the awake section of the session. Then, for each window, the geometric mean of the ratio of the timescales to the awake baseline was computed. After accounting for the baseline awake instability, the degree of destabilization in the timescales in each NHP was very similar. Vertical dotted lines are eyes close/eyes open, with color used to indicate the mean for each NHP.

(C) The distribution of instability values across sessions in each NHP. The x axis is time relative to anesthesia start, and the y axis is the real part of the characteristic roots. Color represents the density of characteristic roots in a particular bin. In both NHPs, the distribution shifted upward to be more unstable during the unconscious state.

(D) The mean instability grouped by section of the session, including at the higher loading dose (0.58 mg/kg/min for NHP 1 and 0.285 mg/kg/min for NHP 2) and the lower maintenance dose (0.32 mg/kg/min for NHP 1 and approx. 0.075 mg/kg/min for NHP 2). For full section definitions, see STAR Methods and Figure S6.

(E) Frequency information from computed roots. The x axis is the real part of the characteristic root, and the y axis is the imaginary part of the characteristic root (converted to a frequency in Hz by dividing by 2π).

See also Figures S2, S5, and S6.

This simple model perfectly captured the qualitative change in neural responses to stimuli from the awake state to the unconscious state. Throughout the entire session, tone and puff stimuli were delivered to the NHPs (see STAR Methods). In our analyses, we focused on "tone-puff" trials consisting of a 500 ms tone at time 0, followed by a 500 ms delay, and then a 10 ms airpuff at time 1. We computed the mean cortical LFP response to the tones and puffs from all sessions (Figures 5D and 5E). There was a phase shift in the sensory responses under anesthesia, and the responses are slower than during the awake state. This exactly matched the intuition from the simulated systems: less stable systems decay back to rest more slowly after being perturbed.

Sensory-evoked trajectories are consistent with destabilized dynamics under anesthesia

We already demonstrated that sensory responses to perturbations are slower in unconsciousness compared with the awake state. We now consider how this destabilization manifests in

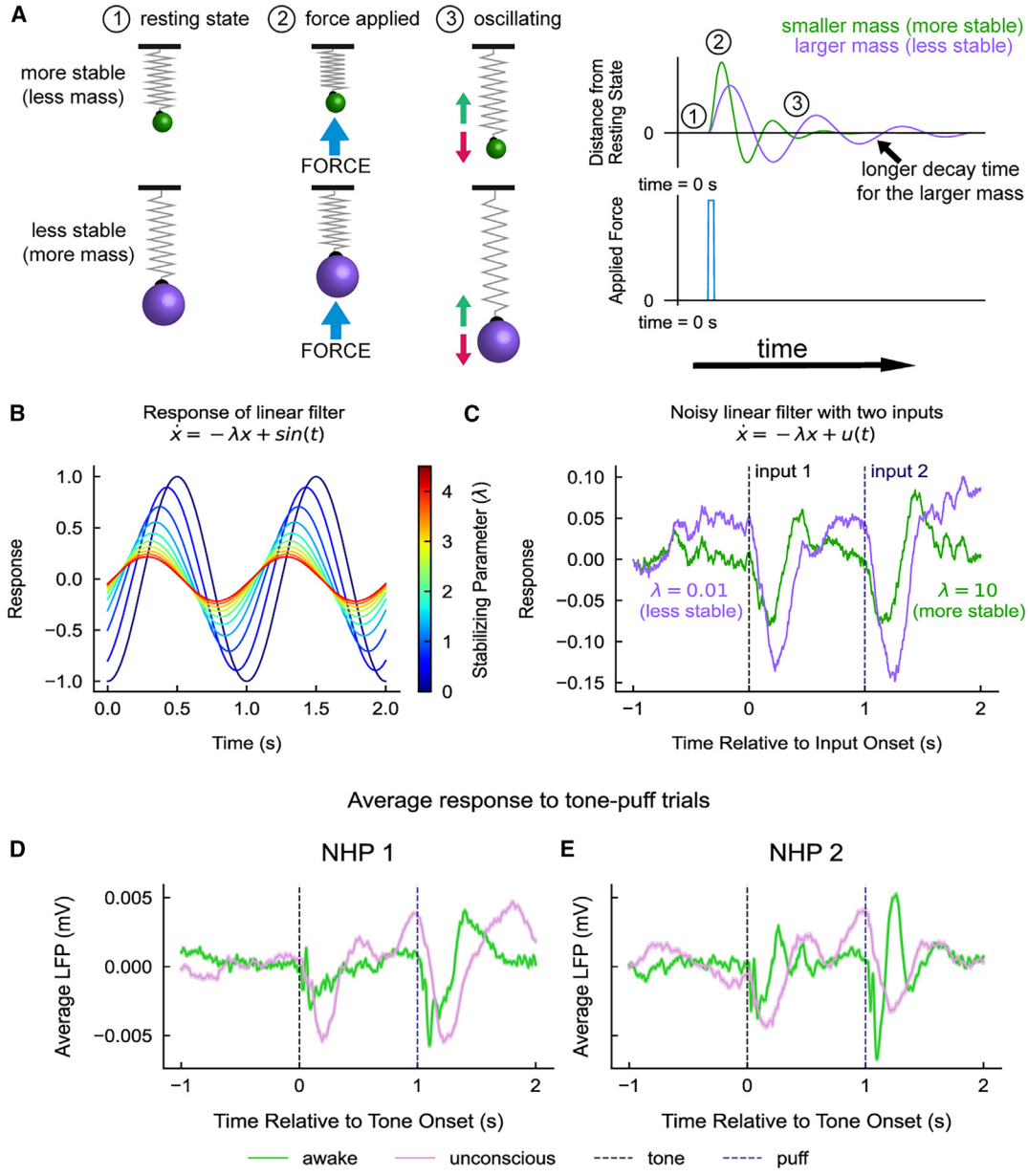


Figure 5. Destabilized dynamics explain sensory responses to tones and puffs during anesthetic unconsciousness

Data in (D) and (E) are represented as mean ± SEM, with results are averaged over trials.

(A) Mass-spring oscillators with spheres of different masses hanging from identical springs. (1) Spheres perturbed upward by equal force (2) oscillate back to the resting state (3). Smaller mass (green) decays faster than larger mass (purple).

(B) Simulations show that increasing filter decay rate—thus stabilizing the filter—generates a phase shift and amplitude decrease for oscillatory input.

(C) Response to two oscillatory inputs and noise for two decay parameters. Less stable (purple) yielded amplitude increase and phase shift relative to more stable (green).

(D and E) Mean tone-evoked LFP response for each NHP (D, NHP 1; E, NHP 2). Anesthetic unconscious state shows a phase shift relative to awake state, matching destabilized linear filters.

the neural state space. To gain some intuition for what we might expect, we consider a conceptual diagram of what happens when stable systems with different timescales of response are perturbed (Figure 6A). In a system with more stability, a perturbation may cause a brief divergence from a baseline region of state space (left). In a system with less stability, however, the perturbation induces a prolonged rotational divergence from the baseline region, depicted with slowly decaying oscillations (right).

We simulated two two-dimensional linear systems with different levels of stability (Figure 6B). At times 0 and 1, the system was perturbed with constant inputs, matching the structure of the tone and puff stimuli presented to the NHPs. As suggested in the conceptual

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

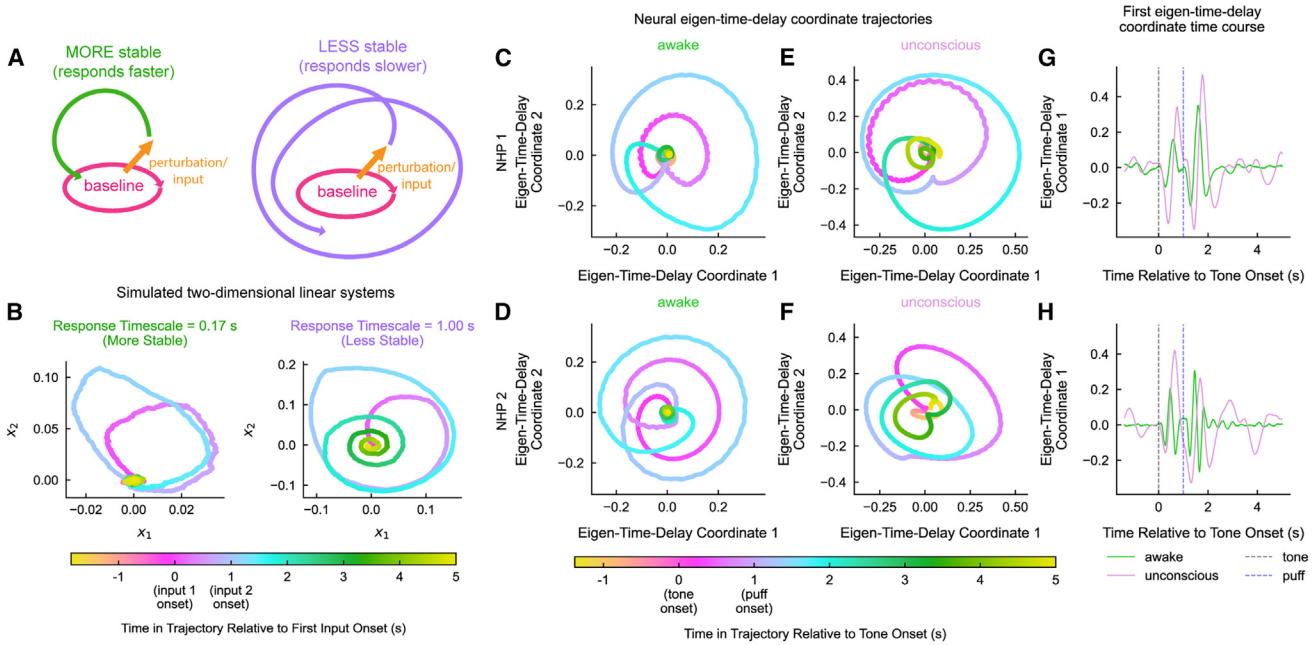


Figure 6. Low-dimensional sensory-evoked trajectories are consistent with destabilized dynamics under anesthesia

(A) Cartoon depiction of dynamic responses to perturbation in stable systems. Without perturbation (orange arrow), the system remains in a small state-space region (pink oval). In more stable systems, there is a quick recovery to baseline after perturbation (left, green arrow); less stable systems show slower recovery, potentially with decaying oscillatory dynamics (right, purple arrow).

(B) Two-dimensional noise-driven linear systems simulated with two different intrinsic timescales. Systems were perturbed with 500 ms input at time 0 and 150 ms input at time 1, approximately matching the stimuli provided to the NHPs. The systems exhibited different responses based on stability (left: more stable, right: less stable).

(C and D) State-space embeddings of the mean responses to tone-puff trials from NHP 1 (C) and NHP 2 (D) during the awake state. Neural responses from tone-puff trials were each delay embedded to illuminate the attractor structure and then averaged. PCA was performed on the mean delay-embedded trajectories for visualization (yielding scaled eigen-time-delay coordinates; see STAR Methods). Awake trajectories rapidly decayed in response to stimuli.

(E and F) Same as (C) and (D), during unconscious state. Trajectories decayed slower with oscillatory structure.

(G and H) First coordinate from awake and unconscious state tone-puff responses for NHPs. The unconscious response was slower.

See also Figure S3.

diagram, the more stable system (shorter governing timescale) was perturbed and then quickly returned back to the baseline. By contrast, the less stable system (longer governing timescale) responded to perturbations with slowly decaying oscillations.

We performed dimensionality reductions for subspace coding on the LFPs from all brain areas collected during both the awake and unconscious states. Our dimensionality reduction approach involved first performing a time-delay embedding on neural data from tone-puff trials from all sessions (we use a delay interval of 20 ms and 32 delays). As discussed, delay embedding can help with attractor reconstruction when the available data constitutes a partial observation from a higher-dimensional system. We then averaged each delay embedding coordinate across trials before performing principal-component analysis (PCA) to obtain a visualization of neural trajectories in two dimensions. This is equivalent to computing (scaled, i.e., non-whitened) eigen-time-delay coordinates.

We visualized two-dimensional state-space trajectories of neural responses to tone-puff trials in eigen-time-delay coordinates (Figures 6C–6F). During the awake state, the state-space trajectories display two clear responses to the two stimuli: the response to the tone, at time 0, followed by the response to the airpuff at time 1 (Figures 6C and 6D). These perturbations caused a deviation from the baseline region into other regions of state space. The system quickly recovered from the perturbation, returning to the baseline state. By contrast, in unconsciousness, the perturbations due to tones and puffs caused prolonged deviations into state space (Figures 6E and 6F). These state-space trajectories are characterized by a slow oscillatory decay back to baseline, as in the simulation. Furthermore, in the awake state, after being first perturbed by the tone, the system was able to decay back to the baseline region before the puff. However, in unconsciousness, the system did not recover from the tone before the puff was delivered. To highlight the different timescales between states, we also visualized the time course of the first eigen-time-delay coordinate (Figures 6G and 6H). The unconscious curve depicts a slower neural response to stimulus perturbation, as suggested by the results from the previous section. The results of this section were consistent across variations in the delay embedding parameters (Figure S3).

Increasing inhibition in random RNNs destabilizes them

We now propose a simple mechanism through which propofol can induce destabilization in neural circuits: increased inhibition.

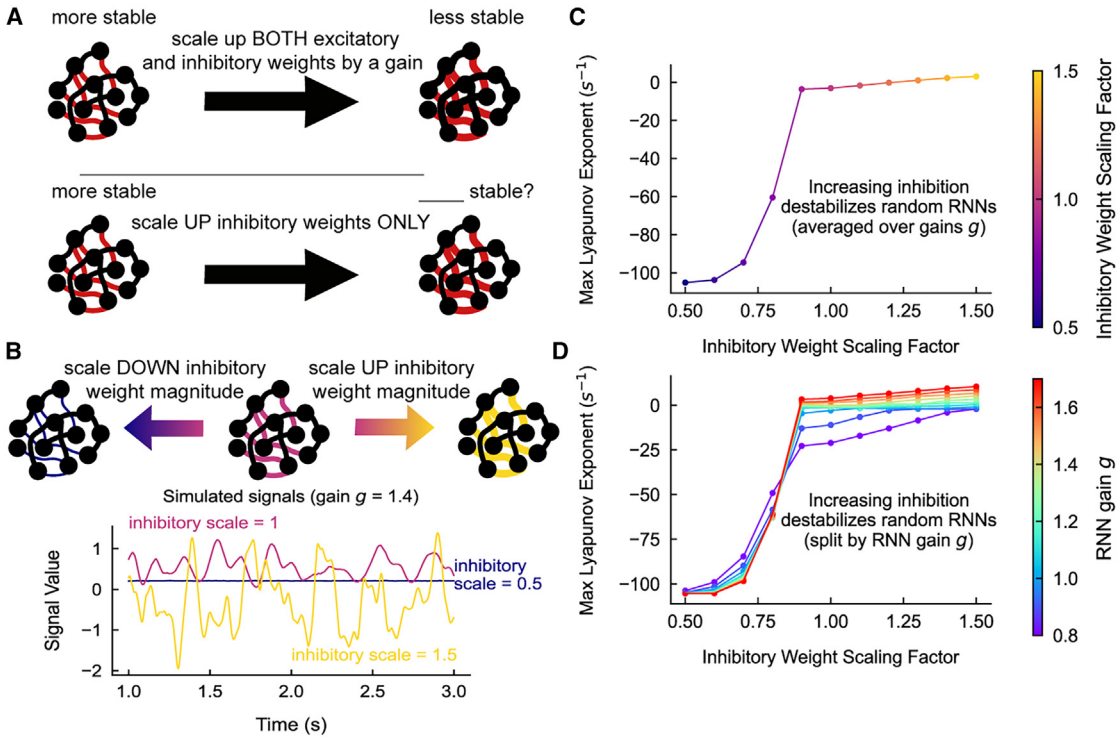


Figure 7. Increasing inhibition in random recurrent neural networks destabilizes them

Data in (C) and (D) are represented as mean ± SEM.

(A) The effect of scaling only inhibitory weights in random RNNs is unknown.

(B) In an RNN driven by noise, scaling up inhibitory weights led to high-amplitude dynamics. Scaling down these weights suppressed amplitudes.

(C) Mean instability (maximum Lyapunov exponent) of RNNs for different scalings of inhibitory weights, averaged over (10) simulations and different baseline gains. Increasing inhibitory weight magnitude destabilized networks, and decreasing the inhibitory weight magnitudes stabilized them.

(D) Same as (C), but split by baseline gain. Increasing inhibitory weights destabilized, while decreasing inhibitory weights stabilized networks.

See also Figure S4.

Propofol is known to act as an agonist at $GABA_A$ inhibitory receptors, thus increasing inhibitory tone. To model the effects of propofol on neural circuits, we alter the connectivity of the randomly connected RNNs (see STAR Methods). These RNNs include a gain parameter g that scales the synaptic weights and induces a transition to chaos in these networks.[97] While it is therefore known that increasing the magnitude of *all* the network weights leads to a destabilization in these networks, the effects of changing *only* the inhibitory weights are unknown (Figures 7A and 7B).

We tested the impact of scaling the magnitude of the inhibitory weights in networks of varying baseline gain (thus impacting the baseline stability of the networks). We simulated the RNN dynamics with a small amount of noise (see STAR Methods). We then measured the network instability through the maximum Lyapunov exponent, computed using standard methods.[95] We found that across all baseline connectivity gains, increasing the inhibitory tone in the networks destabilized them (Figures 7C and 7D). Furthermore, decreasing the inhibitory tone in the networks stabilized them ($p < 0.05$ for all comparisons with one-sided Wilcoxon signed-rank test; see STAR Methods and Figure S4).

We emphasize that this is a somewhat surprising result, given that inhibitory connectivity is often thought of as suppressing activity. In fact, it has been suggested that the increased inhibitory tone during propofol anesthesia might shut off regions of the brain during unconsciousness.[6,99] Nevertheless, the demonstrated destabilization of RNNs through increased inhibitory connectivity lends support to propofol's action at $GABA_A$ receptors underlying the observed destabilization and loss of consciousness during propofol anesthesia.

DISCUSSION

Our results show that propofol anesthesia destabilizes cortical neural dynamics. We found that the stability of brain dynamics was an excellent marker for anesthetic depth, as it smoothly and monotonically varied with the depth of anesthetic state. We then examined neural responses to sensory inputs. We found a longer timescale for recovery from perturbation in unconsciousness, like that seen in destabilized linear systems. We also found that increasing inhibition (as propofol does) in artificial RNN models produced destabilization.

Propofol disrupts the balance between cortical excitation and inhibition. This balance is known to be critical for maintaining the stability of cortical dynamics.[100] Combined with our findings, this paints a picture in which propofol tampers with this balance, causing widespread cortical instabilities and thereby disrupting

the brain's capacity for information processing. Overall, our analysis suggests a mechanism for anesthesia that involves destabilizing brain activity to the point where the brain loses the ability to maintain conscious awareness.

We have also demonstrated the efficacy of a novel approach, DeLASE, designed to directly estimate changes in neural stability from data. The approach brings together multiple aspects of dynamical systems theory, including delay embeddings, Koopman operators, and delay differential equations theory. The linearity of DeLASE enables it to be tractably deployed on ultra-high-volume and high-dimensional neural data. Its theoretically rigorous grounding enables it to provide robust stability estimates despite challenging features of the data such as nonlinearity and partial observation.

The extended response times to stimuli may contribute to the loss of conscious awareness during anesthesia. We discovered that during anesthetic unconsciousness, the response timescales were nearly twice as slow as during the awake state. Therefore, when faced with an input, such as a sensory one, the neural dynamics may not be capable of the synchronization dynamics across areas required to produce conscious awareness. This is consistent with the observation that sensory cortex responses to sensory stimuli are less affected by anesthesia than those in the higher cortex.[101] Though the signal may be present in sensory areas, the full brain—including the higher-level areas thought to be necessary for conscious perception—does not converge to a combined stimulus-guided trajectory fast enough.

We wish to emphasize the use of delay embeddings as part of our methodology for analyzing the nature of neural dynamics across states. Delay embeddings are a widely used tool known to improve the quality of dynamics and attractor reconstruction when observations are of a much smaller dimension than the dimensionality of the true system,[35,36,42,53,54,56–58,61–63,71] as is always the case in neuroscience. Delay embedding the neural data was, in our case, not only beneficial for the estimation of dynamics but also for the visualization of neural trajectories. Patterns in neural response trajectories emerged most clearly when the trajectory history was used to reconstruct the attractor (Figures 6 and S3).

Our findings support the hypothesis that neural computations are instantiated by reliable dynamics in neural state space. Given that computations such as working memory, motor control, and decision-making are instantiated by dynamical features such as fixed points and attractors, destabilizing neural dynamics would undermine the ability of neural circuits to perform these computations.[22,24,27,28,102–108]

While we focused on the comparison between conscious and unconscious states in this paper, we emphasize that our approach to stability estimation was completely agnostic to the nature of the mental state that generated the neural data. It could thus be applied to a wide variety of data from a range of states. In particular, a compelling potential application is to data from psychiatric and mood disorders. Conditions such as depression, anxiety, substance use disorder, and schizophrenia can all be characterized as having distorted thinking patterns relative to neurotypical states, distortions that have been hypothesized to arise from changes to the stability landscape.[19,109–114] Tracking changes in stability in neural dynamics over time for individuals with these conditions could help shape the course of treatment. It could also shed light on the mechanisms of interventions like psychedelics and meditation, which are thought to disrupt overly stable dynamics.[115–117] Many clinical science studies have tracked changes in physiological measures across time—for example, over the course of treatment,[118,119] such as neurofeedback.[120] When considered over the course of treatment, changes in these physiological measures were shown to correspond to decreases in symptom severity. The hypothesized deep connection between neural stability and psychiatric disorders suggests that measuring changes in stability could be an excellent approach for monitoring treatment efficacy.

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.neuron.2024.06.011.

ACKNOWLEDGMENTS

This study was supported by The Office of Naval Research (I.R.F. and N00014-23-1-2768 to E.K.M.), The National Institute of Mental Health 1R01MH131715-01 (E.K.M.), The National Science Foundation Computer and Information Science and Engineering Directorate (I.R.F.), The National Institute of Neurological Disorders and Stroke R01NS123120 (E.N.B.), The Simons Foundation through The Simons Center for the Social Brain (E.K.M.) and The Simons Collaboration on the Global Brain (I.R.F.), The JPB Foundation (E.K.M.), The Picower Institute for Learning and Memory (E.K.M.), and The McGovern Institute at MIT (I.R.F.). We thank Chandrika Prakash Vyasarayani and Antonio Carlos Costa for many helpful discussions and thoughts.

AUTHOR CONTRIBUTIONS

Conceptualization, A.J.E. and L.K.; methodology, A.J.E., L.K., S.C., J.T., and I.R.F.; simulations and data analysis, A.J.E.; macaque surgeries, recordings, and data preprocessing, A.M.B., J.A.D., M.K.M., and S.L.B.; writing – original draft, A.J.E., L.K., I.R.F., and E.K.M.; writing – review & editing, A.J.E., L.K.,

A.M.B., J.A.D., M.K.M., S.L.B., S.C., E.N.B., I.R.F., and E.K.M.; funding acquisition, E.N.B., I.R.F., and E.K.M.; supervision, E.N.B., I.R.F., and E.K.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 31, 2024
Revised: May 13, 2024
Accepted: June 14, 2024
Published: July 15, 2024

REFERENCES

1. Lewis, L.D., Weiner, V.S., Mukamel, E.A., Donoghue, J.A., Eskandar, E.N., Madsen, J.R., Anderson, W.S., Hochberg, L.R., Cash, S.S., Brown, E.N., et al. (2012). Rapid fragmentation of neuronal networks at the onset of propofol-induced unconsciousness. Proc. Natl. Acad. Sci. USA *109*, E3377–E3386. https://doi.org/10.1073/pnas.1210907109.
2. Bastos, A.M., Donoghue, J.A., Brincat, S.L., Mahnke, M., Yanar, J., Correa, J., Waite, A.S., Lundqvist, M., Roy, J., Brown, E.N., et al. (2021). Neural effects of propofol-induced unconsciousness and its reversal using thalamic stimulation. eLife *10*, e60824. https://doi.org/10.7554/eLife.60824.
3. Flores, F.J., Hartnack, K.E., Fath, A.B., Kim, S.-E., Wilson, M.A., Brown, E.N., and Purdon, P.L. (2017). Thalamocortical synchronization during induction and emergence from propofol-induced unconsciousness. Proc. Natl. Acad. Sci. USA *114*, E6660–E6668. https://doi.org/10.1073/pnas.1700148114.
4. Palva, S., and Palva, J.M. (2007). New vistas for alpha-frequency band oscillations. Trends Neurosci. *30*, 150–158. https://doi.org/10.1016/j.tins.2007.02.001.
5. Ching, S., Cimenser, A., Purdon, P.L., Brown, E.N., and Kopell, N.J. (2010). Thalamocortical model for a propofol-induced alpha-rhythm associated with loss of consciousness. Proc. Natl. Acad. Sci. USA. *107*, 22665–22670. https://doi.org/10.1073/pnas.1017069108.
6. Brown, E.N., Lydic, R., and Schiff, N.D. (2010). General anesthesia, sleep, and coma. N. Engl. J. Med. *363*, 2638–2650. https://doi.org/10.1056/NEJMra0808281.
7. Saalmann, Y.B., and Kastner, S. (2015). The cognitive thalamus. Front. Syst. Neurosci. *9*, 39. https://doi.org/10.3389/fnsys.2015.00039.
8. Seth, A.K., and Bayne, T. (2022). Theories of consciousness. Nat. Rev. Neurosci. *23*, 439–452. https://doi.org/10.1038/s41583-022-00587-4.
9. Baars, B.J. (1988). A Cognitive Theory of Consciousness (Cambridge University Press).
10. Dehaene, S., and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. Neuron *70*, 200–227. https://doi.org/10.1016/j.neuron.2011.03.018.
11. Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. Biol. Bull. *215*, 216–242. https://doi.org/10.2307/25470707.
12. Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. Nat. Rev. Neurosci. *17*, 450–461. https://doi.org/10.1038/nrn.2016.44.
13. Graziano, M.S.A. (2017). The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. Front. Robot. AI *4*. https://doi.org/10.3389/frobt.2017.00060.
14. Brown, R., Lau, H., and LeDoux, J.E. (2019). Understanding the Higher-Order Approach to Consciousness. Trends Cogn. Sci. *23*, 754–768. https://doi.org/10.1016/j.tics.2019.06.009.
15. Dehaene, S., Sergent, C., and Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc. Natl. Acad. Sci. USA *100*, 8520–8525. https://doi.org/10.1073/pnas.1332574100.
16. Mashour, G.A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron *105*, 776–798. https://doi.org/10.1016/j.neuron.2020.01.026.
17. Babloyantz, A., and Destexhe, A. (1986). Low-dimensional chaos in an instance of epilepsy. Proc. Natl. Acad. Sci. USA *83*, 3513–3517. https://doi.org/10.1073/pnas.83.10.3513.
18. Theiler, J. (1995). On the evidence for low-dimensional chaos in an epileptic electroencephalogram. Phys. Lett. A *196*, 335–341. https://doi.org/10.1016/0375-9601(94)00856-K.
19. Carhart-Harris, R.L., Leech, R., Hellyer, P.J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D.R., and Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. Front. Hum. Neurosci. *8*, 20. https://doi.org/10.3389/fnhum.2014.00020.
20. Beggs, J.M. (2008). The criticality hypothesis: how local cortical networks might optimize information processing. Philos. Trans. A Math. Phys. Eng. Sci. *366*, 329–343. https://doi.org/10.1098/rsta.2007.2092.
21. Kozachkov, L., Lundqvist, M., Slotine, J.J., and Miller, E.K. (2020). Achieving stable dynamics in neural circuits. PLoS Comput. Biol. *16*, e1007659. https://doi.org/10.1371/journal.pcbi.1007659.
22. Vyas, S., Golub, M.D., Sussillo, D., and Shenoy, K.V. (2020). Computation Through Neural Population Dynamics. Annu. Rev. Neurosci. *43*, 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115.
23. Demertzi, A., Tagliazucchi, E., Dehaene, S., Deco, G., Barttfeld, P., Raimondo, F., Martial, C., Fernández-Espejo, D., Rohaut, B., Voss, H.U., et al. (2019). Human consciousness is supported by dynamic complex patterns of brain signal coordination. Sci. Adv. *5*, eaat7603. https://doi.org/10.1126/sciadv.aat7603.
24. Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA *79*, 2554–2558. https://doi.org/10.1073/pnas.79.8.2554.
25. Hirsch, M.W. (1989). Convergent activation dynamics in continuous time networks. Neural Netw. *2*, 331–349. https://doi.org/10.1016/0893-6080(89)90018-X.
26. Cohen, M.A., and Grossberg, S. (1987). Absolute Stability of Global Pattern Formation and Parallel Memory Storage by Competitive Neural Networks. In The Adaptive Brain I – Cognition, Learning, Reinforcement, and Rhythm, S. Grossberg, ed. (North-Holland), pp. 288–308. https://doi.org/10.1016/S0166-4115(08)60913-9.
27. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56. https://doi.org/10.1038/nature11129.
28. Sohn, H., Narain, D., Meirhaeghe, N., and Jazayeri, M. (2019). Bayesian Computation through Cortical Latent Dynamics. Neuron *103*, 934–947.e5. https://doi.org/10.1016/j.neuron.2019.06.012.
29. Lorenz, E.N. (1963). Deterministic Nonperiodic Flow. J. Atmos. Sci. *20*, 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.
30. Toker, D., Pappas, I., Lendner, J.D., Frohlich, J., Mateos, D.M., Muthukumaraswamy, S., Carhart-Harris, R., Paff, M., Vespa, P.M., Monti, M.M., et al. (2022). Consciousness is supported by near-critical slow cortical electrodynamics. Proc. Natl. Acad. Sci. USA *119*, e2024455119. https://doi.org/10.1073/pnas.2024455119.
31. López-González, A., Panda, R., Ponce-Alvarez, A., Zamora-López, G., Escrichs, A., Martial, C., Thibaut, A., Gosseries, O., Kringelbach, M.L., Annen, J., et al. (2021). Loss of consciousness reduces the stability of brain hubs and the heterogeneity of brain dynamics. Commun. Biol. *4*, 1037. https://doi.org/10.1038/s42003-021-02537-9.
32. Solovey, G., Alonso, L.M., Yanagawa, T., Fujii, N., Magnasco, M.O., Cecchi, G.A., and Proekt, A. (2015). Loss of Consciousness Is Associated with Stabilization of Cortical Activity. J. Neurosci. *35*, 10866–10877. https://doi.org/10.1523/JNEUROSCI.4895-14.2015.

33. Alonso, L.M., Proekt, A., Schwartz, T.H., Pryor, K.O., Cecchi, G.A., and Magnasco, M.O. (2014). Dynamical criticality during induction of anesthesia in human ECoG recordings. Front. Neural Circuits *8*, 20. https://doi.org/10.3389/fncir.2014.00020.

34. Krzemiński, D., Kamiński, M., Marchewka, A., and Bola, M. (2017). Breakdown of long-range temporal correlations in brain oscillations during general anesthesia. Neuroimage *159*, 146–158. https://doi.org/10.1016/j.neuroimage.2017.07.047.

35. Takens, F. (1981). Detecting strange attractors in turbulence. In Dynamical Systems and Turbulence, Warwick 1980 (Springer), pp. 366–381. https://doi.org/10.1007/BFb0091924.

36. 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.

37. Kamb, M., Kaiser, E., Brunton, S.L., and Kutz, J.N. (2020). Time-Delay Observables for Koopman: Theory and Applications. SIAM J. Appl. Dyn. Syst. *19*, 886–917. https://doi.org/10.1137/18M1216572.

38. Costa, A.C., Ahamed, T., Jordan, D., and Stephens, G.J. (2023). Maximally predictive states: From partial observations to long timescales. Chaos *33*, 23136. https://doi.org/10.1063/5.0129398.

39. Dhir, N., Kosiorek, A.R., and Posner, I. (2017). Bayesian delay embeddings for dynamical systems. In 31st Conference on Neural Information Processing Systems (NIPS 2017) https://www.robots.ox.ac.uk/~mobile/Papers/2017NIPS_dhir.pdf.

40. Susuki, Y., and Mezić, I. (2015). A prony approximation of Koopman Mode Decomposition. In 2015 54th IEEE Conference on Decision and Control (CDC), pp. 7022–7027. https://doi.org/10.1109/CDC.2015.7403326.

41. Arbabi, H., Korda, M., and Mezić, I. (2018). A Data-Driven Koopman Model Predictive Control Framework for Nonlinear Partial Differential Equations. In IEEE Conference on Decision and Control. (CDC), pp. 6409–6414. https://doi.org/10.1109/CDC.2018.8619720.

42. Brunton, B.W., Johnson, L.A., Ojemann, J.G., and Kutz, J.N. (2016). Extracting spatial-temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. J. Neurosci. Methods *258*, 1–15. https://doi.org/10.1016/j.jneumeth.2015.10.010.

43. Bakarji, J., Champion, K., Nathan Kutz, J., and Brunton, S.L. (2023). Discovering governing equations from partial measurements with deep delay autoencoders. Proc. R. Soc. A. *479*, 20230422. https://doi.org/10.1098/rspa.2023.0422.

44. Axås, J., and Haller, G. (2023). Model reduction for nonlinearizable dynamics via delay-embedded spectral submanifolds. Nonlinear Dyn. *111*, 22079–22099. https://doi.org/10.1007/s11071-023-08705-2.

45. Juang, J.N., and Pappa, R.S. (1985). An Eigensystem Realization Algorithm (ERA) for modal parameter identification and model reduction. In JPL Proc. of the Workshop on Identification and Control of Flexible Space Struct., *3*.

46. Khodkar, M.A., and Hassanzadeh, P. (2021). A data-driven, physics-informed framework for forecasting the spatiotemporal evolution of chaotic dynamics with nonlinearities modeled as exogenous forcings. J. Comput. Phys. *440*, 110412. https://doi.org/10.1016/j.jcp.2021.110412.

47. Gauthier, D.J., Bollt, E., Griffith, A., and Barbosa, W.A.S. (2021). Next generation reservoir computing. Nat. Commun. *12*, 5564. https://doi.org/10.1038/s41467-021-25801-2.

48. Crutchfield, J., and McNamara, B.S. (1987). Equations of Motion from a Data Series. Complex Syst. *1*, 417–452.

49. Sugihara, G., and May, R.M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature *344*, 734–741. https://doi.org/10.1038/344734a0.

50. Rowlands, G., and Sprott, J.C. (1992). Extraction of dynamical equations from chaotic data. Phys. D *58*, 251–259. https://doi.org/10.1016/0167-2789(92)90113-2.

51. Abarbanel, H.D.I., Brown, R., Sidorowich, J.J., and Tsimring, L.S. (1993). The analysis of observed chaotic data in physical systems. Rev. Mod. Phys. *65*, 1331–1392. https://doi.org/10.1103/RevModPhys.65.1331.

52. Ye, H., Deyle, E.R., Gilarranz, L.J., and Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. Sci. Rep. *5*, 14750. https://doi.org/10.1038/srep14750.

53. Ye, H., Beamish, R.J., Glaser, S.M., Grant, S.C.H., Hsieh, C.-H., Richards, L.J., Schnute, J.T., and Sugihara, G. (2015). Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. Proc. Natl. Acad. Sci. USA *112*, E1569–E1576. https://doi.org/10.1073/pnas.1417063112.

54. Deyle, E.R., May, R.M., Munch, S.B., and Sugihara, G. (2016). Tracking and forecasting ecosystem interactions in real time. Proc. Biol. Sci. *283*, 20152258. https://doi.org/10.1098/rspb.2015.2258.

55. Park, J., Pao, G.M., Sugihara, G., Stabenau, E., and Lorimer, T. (2022). Empirical mode modeling. Nonlinear Dyn. *108*, 2147–2160. https://doi.org/10.1007/s11071-022-07311-y.

56. Tajima, S., Yanagawa, T., Fujii, N., and Toyoizumi, T. (2015). Untangling Brain-Wide Dynamics in Consciousness by Cross-Embedding. PLoS Comput. Biol. *11*, e1004537. https://doi.org/10.1371/journal.pcbi.1004537.

57. Ahamed, T., Costa, A.C., and Stephens, G.J. (2021). Capturing the continuous complexity of behaviour in Caenorhabditis elegans. Nat. Phys. *17*, 275–283. https://doi.org/10.1038/s41567-020-01036-8.

58. Gilpin, W. (2020). Deep reconstruction of strange attractors from time series. In Proceedings of the 34th International Conference on Neural Information Processing Systems NIPS'20 (Curran Associates Inc.), pp. 204–216.

59. Watanakeesuntorn, W., Takahashi, K., Ichikawa, K., Park, J., Sugihara, G., Takano, R., Haga, J., and Pao, G.M. (2020). Massively Parallel Causal Inference of Whole Brain Dynamics at Single Neuron Resolution. In 2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS) (IEEE) (IEEE Publications), pp. 196–205. https://doi.org/10.1109/ICPADS51040.2020.00035.

60. Zhang, Z., Li, K., and Hu, X. (2023). Mapping nonlinear brain dynamics by phase space embedding with fMRI data. Biomed. Signal Process. Control *82*, 104521. https://doi.org/10.1016/j.bspc.2022.104521.

61. Raut, R.V., Rosenthal, Z.P., Wang, X., Miao, H., Zhang, Z., Lee, J.-M., Raichle, M.E., Bauer, A.Q., Brunton, S.L., Brunton, B.W., et al. (2023). Arousal as a universal embedding for spatiotemporal brain dynamics. Preprint at bioRxiv. https://doi.org/10.1101/2023.11.06.565918.

62. Mori, H. (1965). Transport, Collective Motion, and Brownian Motion. Prog. Theor. Phys. *33*, 423–455. https://doi.org/10.1143/PTP.33.423.

63. Zwanzig, R. (1973). Nonlinear generalized Langevin equations. J. Stat. Phys. *9*, 215–220. https://doi.org/10.1007/BF01008729.

64. Lin, Y.T., Tian, Y., Livescu, D., and Anghel, M. (2021). Data-Driven Learning for the Mori–Zwanzig Formalism: A Generalization of the Koopman Learning Framework. SIAM J. Appl. Dyn. Syst. *20*, 2558–2601. https://doi.org/10.1137/21M1401759.

65. Chorin, A.J., Hald, O.H., and Kupferman, R. (2000). Optimal prediction and the Mori-Zwanzig representation of irreversible processes. Proc. Natl. Acad. Sci. USA *97*, 2968–2973. https://doi.org/10.1073/pnas.97.7.2968.

66. Lin, K.K., and Lu, F. (2021). Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism. J. Comput. Phys. *424*, 109864. https://doi.org/10.1016/j.jcp.2020.109864.

67. Hijón, C., Español, P., Vanden-Eijnden, E., and Delgado-Buscalioni, R. (2010). Mori-Zwanzig formalism as a practical computational tool. discussion 323–345, 467–481. Faraday Discuss. *144*, 301–322. https://doi.org/10.1039/b902479b.

68. Koopman, B.O. (1931). Hamiltonian Systems and Transformation in Hilbert Space. Proc. Natl. Acad. Sci. USA *17*, 315–318. https://doi.org/10.1073/pnas.17.5.315.

69. 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.

70. Mezić, I. (2005). Spectral Properties of Dynamical Systems, Model Reduction and Decompositions. Nonlinear Dyn. *41*, 309–325. https://doi.org/10.1007/s11071-005-2824-x.

71. Arbabi, H., and Mezić, I. (2017). Ergodic Theory, Dynamic Mode Decomposition, and Computation of Spectral Properties of the Koopman Operator. SIAM J. Appl. Dyn. Syst. *16*, 2096–2126. https://doi.org/10.1137/17M1125236.

72. Kusaba, A., Kuboyama, T., Shin, K., Sasaki, M., and Inagaki, S. (2022). A new combination of Hankel and sparsity-promoting dynamic mode decompositions and its application to the prediction of plasma turbulence. Jpn. J. Appl. Phys. *61*, SA1011. https://doi.org/10.35848/1347-4065/ac1c3c.

73. Hirsh, S.M., Ichinaga, S.M., Brunton, S.L., Nathan Kutz, J., and Brunton, B.W. (2021). Structured time-delay models for dynamical systems with connections to Frenet-Serret frame. Proc. Math. Phys. Eng. Sci. *477*, 20210097. https://doi.org/10.1098/rspa.2021.0097.

74. Schmid, P.J. (2010). Dynamic mode decomposition of numerical and experimental data. J. Fluid Mech. *656*, 5–28. https://doi.org/10.1017/S0022112010001217.

75. Rowley, C.W., Mezić, I., Bagheri, S., Schlatter, P., and Henningson, D.S. (2009). Spectral analysis of nonlinear flows. J. Fluid Mech. *641*, 115–127. https://doi.org/10.1017/S0022112009992059.

76. Tu, J.H., Rowley, C.W., Nathan Kutz, J., Brunton, S.L., and Nathan Kutz, J. (2014). On dynamic mode decomposition: Theory and applications. J. Comput. Dyn. *1*, 391–421. https://doi.org/10.3934/jcd.2014.1.391.

77. Williams, M.O., Kevrekidis, I.G., and Rowley, C.W. (2015). A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. J. Nonlinear Sci. *25*, 1307–1346. https://doi.org/10.1007/s00332-015-9258-5.

78. Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. USA *113*, 3932–3937. https://doi.org/10.1073/pnas.1517384113.

79. Williams, M.O., Rowley, C.W., and Kevrekidis, I.G. (2015). A kernel-based method for data-driven koopman spectral analysis. J. Comp. Dyn. *2*, 247–265.

80. Folkestad, C., Pastor, D., Mezic, I., Mohr, R., Fonoberova, M., and Burdick, J. (2020). Extended Dynamic Mode Decomposition with Learned Koopman Eigenfunctions for Prediction and Control. In 2020 American Control Conference (ACC) (IEEE), pp. 3906–3913. https://doi.org/10.23919/ACC45564.2020.9147729.

81. Alford-Lago, D.J., Curtis, C.W., Ihler, A.T., and Issan, O. (2022). Deep learning enhanced dynamic mode decomposition. Chaos *32*, 33116. https://doi.org/10.1063/5.0073893.

82. Lusch, B., Kutz, J.N., and Brunton, S.L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. Nat. Commun. *9*, 4950. https://doi.org/10.1038/s41467-018-07210-0.

83. Takeishi, N., Kawahara, Y., and Yairi, T. (2017). Learning Koopman invariant subspaces for dynamic mode decomposition. In Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17 (Curran Associates Inc.), pp. 1130–1140.

84. Kutz, J.N., Fu, X., and Brunton, S.L. (2016). Multiresolution Dynamic Mode Decomposition. SIAM J. Appl. Dyn. Syst. *15*, 713–735. https://doi.org/10.1137/15M1023543.

85. Brunton, S.L., Brunton, B.W., Proctor, J.L., and Kutz, J.N. (2016). Koopman Invariant Subspaces and Finite Linear Representations of Nonlinear Dynamical Systems for Control. PLOS One *11*, e0150171. https://doi.org/10.1371/journal.pone.0150171.

86. Ferre, J., Rokem, A., Buffalo, E.A., Nathan Kutz, J., and Fairhall, A. (2023). Non-Stationary Dynamic Mode Decomposition. IEEE Access *11*, 117159–117176. https://doi.org/10.1101/2023.08.08.552333.

87. Solaija, M.S.J., Saleem, S., Khurshid, K., Hassan, S.A., and Kamboh, A.M. (2018). Dynamic Mode Decomposition Based Epileptic Seizure Detection from Scalp EEG. IEEE Access *6*, 38683–38692. https://doi.org/10.1109/ACCESS.2018.2853125.

88. Bilal, M., Rizwan, M., Saleem, S., Khan, M.M., Alkatheir, M.S., and Alqarni, M. (2019). Automatic Seizure Detection Using Multi-Resolution Dynamic Mode Decomposition. IEEE Access *7*, 61180–61194. https://doi.org/10.1109/ACCESS.2019.2915609.

89. Ostrow, M., Eisen, A.J., Kozachkov, L., and Fiete, I.R. (2023). Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. In Advances in Neural Information Processing Systems proceedings.neurips.cc.

90. Marrouch, N., Slawinska, J., Giannakis, D., and Read, H.L. (2020). Data-driven Koopman operator approach for computational neuroscience. Ann. Math. Artif. Intell. *88*, 1155–1173. https://doi.org/10.1007/s10472-019-09666-2.

91. van der Pol, B. (1926). LXXXVIII. On "relaxation-oscillations.". Lond. Edinb. Dublin Philos. Mag. J. Sci. *2*, 978–992.

92. Verheyden, K., Luzyanina, T., and Roose, D. (2008). Efficient computation of characteristic roots of delay differential equations using LMS methods. J. Comput. Appl. Math. *214*, 209–226. https://doi.org/10.1016/j.cam.2007.02.025.

93. Breda, D., Maset, S., and Vermiglio, R. (2009). TRACE-DDE: a Tool for Robust Analysis and Characteristic Equations for Delay Differential Equations. In Topics in Time Delay Systems: Analysis, Algorithms and Control, J.J. Loiseau, W. Michiels, S.-I. Niculescu, and R. Sipahi, eds. (Springer), pp. 145–155. https://doi.org/10.1007/978-3-642-02897-7_13.

94. Chialvo, D.R. (2010). Emergent complex neural dynamics. Nat. Phys. *6*, 744–750. https://doi.org/10.1038/nphys1803.

95. Dieci, L., Russell, R.D., and Van Vleck, E.S. (1997). On the Computation of Lyapunov Exponents for Continuous Dynamical Systems. SIAM J. Numer. Anal. *34*, 402–423. https://doi.org/10.1137/S0036142993247311.

96. Mohan, A., Vijesh, V., Thumba, D.A., and Kumar, K.S. (2020). Recurrence Network-Based Approach to Distinguish Between Chaotic and Quasiperiodic Solution. In Advances in Signal Processing and Intelligent Recognition Systems (Springer), pp. 368–375. https://doi.org/10.1007/978-981-15-4828-4_30.

97. Sompolinsky, H., Crisanti, A., and Sommers, H.J. (1988). Chaos in random neural networks. Phys. Rev. Lett. *61*, 259–262. https://doi.org/10.1103/PhysRevLett.61.259.

98. Solovey, G., Miller, K.J., Ojemann, J.G., Magnasco, M.O., and Cecchi, G.A. (2012). Self-Regulated Dynamical Criticality in Human ECoG. Front. Integr. Neurosci. *6*, 44. https://doi.org/10.3389/fnint.2012.00044.

99. Bai, D., Pennefather, P.S., MacDonald, J.F., and Orser, B.A. (1999). The general anesthetic propofol slows deactivation and desensitization of GABA(A) receptors. J. Neurosci. *19*, 10635–10646. https://doi.org/10.1523/JNEUROSCI.19-24-10635.1999.

100. Murphy, B.K., and Miller, K.D. (2009). Balanced amplification: a new mechanism of selective amplification of neural activity patterns. Neuron *61*, 635–648. https://doi.org/10.1016/j.neuron.2009.02.005.

101. Tauber, J.M., Brincat, S.L., Stephen, E.P., Donoghue, J.A., Kozachkov, L., Brown, E.N., and Miller, E.K. (2024). Propofol-mediated Unconsciousness Disrupts Progression of Sensory Signals through the Cortical Hierarchy. J. Cogn. Neurosci. *36*, 394–413.

102. Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic Theory of Working Memory. Science *319*, 1543–1546. https://doi.org/10.1126/science.1150769.

103. Churchland, A.K., Kiani, R., and Shadlen, M.N. (2008). Decision-making with multiple alternatives. Nat. Neurosci. *11*, 693–702. https://doi.org/10.1038/nn.2123.

104. Khona, M., and Fiete, I.R. (2022). Attractor and integrator networks in the brain. Nat. Rev. Neurosci. *23*, 744–766. https://doi.org/10.1038/s41583-022-00642-0.

105. Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., and Fiete, I. (2019). The intrinsic attractor manifold and population dynamics of a canonical

cognitive circuit across waking and sleep. Nat. Neurosci. *22*, 1512–1520. https://doi.org/10.1038/s41593-019-0460-x.

106. Mastrogiuseppe, F., and Ostojic, S. (2018). Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. Neuron *99*, 609–623.e29. https://doi.org/10.1016/j.neuron.2018.07.003.
107. Hasselmo, M.E., and Brandon, M.P. (2012). A model combining oscillations and attractor dynamics for generation of grid cell firing. Front. Neural Circuits *6*, 30. https://doi.org/10.3389/fncir.2012.00030.
108. Libby, A., and Buschman, T.J. (2021). Rotational dynamics reduce interference between sensory and memory representations. Nat. Neurosci. *24*, 715–726. https://doi.org/10.1038/s41593-021-00821-9.
109. Holtzheimer, P.E., and Mayberg, H.S. (2011). Stuck in a rut: rethinking depression and its treatment. Trends Neurosci. *34*, 1–9. https://doi.org/10.1016/j.tins.2010.10.004.
110. Zullino, D.F., and Khazaal, Y. (2008). The "rut metaphor": a conceptualization of attractor-shaping properties of addictive drugs. Subst. Use Misuse *43*, 469–479. https://doi.org/10.1080/10826080701205042.
111. Carhart-Harris, R.L., Chandaria, S., Erritzoe, D.E., Gazzaley, A., Girn, M., Kettner, H., Mediano, P.A.M., Nutt, D.J., Rosas, F.E., Roseman, L., et al. (2023). Canalization and plasticity in psychopathology. Neuropharmacology *226*, 109398. https://doi.org/10.1016/j.neuropharm.2022.109398.
112. Juliani, A., Safron, A., and Kanai, R. (2023). Deep CANALs: A deep learning approach to refining the canalization theory of psychopathology. Neurosci. Conscious. *2024*, niae005. https://doi.org/10.31234/osf.io/uxmz6.
113. Braun, U., Harneit, A., Pergola, G., Menara, T., Schäfer, A., Betzel, R.F., Zang, Z., Schweiger, J.I., Zhang, X., Schwarz, K., et al. (2021). Brain network dynamics during working memory are modulated by dopamine and diminished in schizophrenia. Nat. Commun. *12*, 3478. https://doi.org/10.1038/s41467-021-23694-9.
114. Mahadevan, A.S., Cornblath, E.J., Lydon-Staley, D.M., Zhou, D., Parkes, L., Larsen, B., Adebimpe, A., Kahn, A.E., Gur, R.C., Gur, R.E., et al. (2023). Alprazolam modulates persistence energy during emotion processing in first-degree relatives of individuals with schizophrenia: a network control study. Mol. Psychiatry *28*, 3314–3323. https://doi.org/10.1038/s41380-023-02121-z.
115. Singleton, S.P., Luppi, A.I., Carhart-Harris, R.L., Cruzat, J., Roseman, L., Nutt, D.J., Deco, G., Kringelbach, M.L., Stamatakis, E.A., and Kuceyeski, A. (2022). Receptor-informed network control theory links LSD and psilocybin to a flattening of the brain's control energy landscape. Nat. Commun. *13*, 5812. https://doi.org/10.1038/s41467-022-33578-1.
116. Ruffini, G., Lopez-Sola, E., Vohryzek, J., and Sanchez-Todo, R. (2023). Neural geometrodynamics: a psychedelic perspective. Preprint at bioRxiv. https://doi.org/10.1101/2023.08.14.553258.
117. Zhou, D., Kang, Y., Cosme, D., Jovanova, M., He, X., Mahadevan, A., Ahn, J., Stanoi, O., Brynildsen, J.K., Cooper, N., et al. (2023). Mindful attention promotes control of brain network dynamics for self-regulation and discontinues the past from the present. Proc. Natl. Acad. Sci. USA *120*, e2201074119. https://doi.org/10.1073/pnas.2201074119.
118. Helpman, L., Marin, M.-F., Papini, S., Zhu, X., Sullivan, G.M., Schneier, F., Neria, M., Shvil, E., Malaga Aragon, M.J., Markowitz, J.C., et al. (2016). Neural changes in extinction recall following prolonged exposure treatment for PTSD: A longitudinal fMRI study. NeuroImage Clin. *12*, 715–723. https://doi.org/10.1016/j.nicl.2016.10.007.
119. Tan, G., Dao, T.K., Farmer, L., Sutherland, R.J., and Gevirtz, R. (2011). Heart rate variability (HRV) and posttraumatic stress disorder (PTSD): a pilot study. Appl. Psychophysiol. Biofeedback *36*, 27–35. https://doi.org/10.1007/s10484-010-9141-y.
120. Compère, L., Siegle, G.J., Lazzaro, S., Riley, E., Strege, M., Canovali, G., Barb, S., Huppert, T., and Young, K. (2024). Amygdala real-time fMRI neurofeedback upregulation in treatment resistant depression: Proof of concept and dose determination. Behav. Res. Ther. *176*, 104523. https://doi.org/10.1016/j.brat.2024.104523.
121. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.
122. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (SciPy), pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.
123. 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. In NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 8026–8037.
124. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
125. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
126. Dickey, A.S., Suminski, A., Amit, Y., and Hatsopoulos, N.G. (2009). Single-unit stability using chronically implanted multielectrode arrays. J. Neurophysiol. *102*, 1331–1339. https://doi.org/10.1152/jn.90920.2008.
127. Solé-Casals, J., and Vialatte, F.-B. (2015). Towards Semi-Automatic Artifact Rejection for the Improvement of Alzheimer's Disease Screening from EEG Signals. Sensors (Basel) *15*, 17963–17976. https://doi.org/10.3390/s150817963.
128. Urigüen, J.A., and Garcia-Zapirain, B. (2015). EEG artifact removal-state-of-the-art and guidelines. J. Neural Eng. *12*, 031001. https://doi.org/10.1088/1741-2560/12/3/031001.
129. Muthukumaraswamy, S.D. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. Front. Hum. Neurosci. *7*, 138. https://doi.org/10.3389/fnhum.2013.00138.
130. Brunner, D.P., Vasko, R.C., Detka, C.S., Monahan, J.P., Reynolds, C.F., 3rd, and Kupfer, D.J. (1996). Muscle artifacts in the sleep EEG: automated detection and effect on all-night EEG power spectra. J. Sleep Res. *5*, 155–164. https://doi.org/10.1046/j.1365-2869.1996.00009.x.
131. Breda, D. (2012). On characteristic roots and stability charts of delay differential equations. Int. J. Robust Nonlinear Control *22*, 892–917. https://doi.org/10.1002/rnc.1734.
132. Conway, J.B. (1978). Functions of One Complex Variable I (Springer).
133. Ai, B., Sentis, L., Paine, N., Han, S., Mok, A., and Fok, C.-L. (2016). Stability and Performance Analysis of Time-Delayed Actuator Control Systems. J. Dyn. Syst. Meas. Control *138*, 51005. https://doi.org/10.1115/1.4032461.
134. Wahi, P., and Chatterjee, A. (2003). Galerkin Projections for Delay Differential Equations. In ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (American Society of Mechanical Engineers), pp. 2211–2220. https://doi.org/10.1115/DETC2003/VIB-48570.
135. Vyasarayani, C.P., Subhash, S., and Kalmár-Nagy, T. (2014). Spectral approximations for characteristic roots of delay differential equations. Int. J. Dyn. Control *2*, 126–132. https://doi.org/10.1007/s40435-014-0060-2.
136. Hill, S.A. (2004). Pharmacokinetics of drug infusions. Contin. Educ. Anaesth. Crit. Care Pain *4*, 76–80. https://doi.org/10.1093/bjaceaccp/mkh021.
137. Schüttler, J., and Ihmsen, H. (2000). Population pharmacokinetics of propofol: a multicenter study. Anesthesiology *92*, 727–738. https://doi.org/10.1097/00000542-200003000-00017.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
Rhesus macaque (*Macaca mulatta*)	Alpha Genesis Inc. (NHP 1) and Merck & Co., Inc. (NHP 2)	N/A
Software and algorithms		
Python	Python Software Foundation	https://www.python.org/
Code used to analyze electrophysiology data, simulate, and generate figures	This paper	https://github.com/adamjeisen/ChaoticConsciousness https://doi.org/10.5281/zenodo.11167882
Code used to perform the stability analysis (DeLASE)	This paper	https://github.com/adamjeisen/DeLASE https://doi.org/10.5281/zenodo.11168144
Hydra	Meta Platforms, Inc	https://hydra.cc/
Matplotlib	The Matplotlib development team; Hunter[121]	https://matplotlib.org/stable/ https://doi.org/10.5281/zenodo.10059757
Pandas	NumFOCUS, Inc.; McKinney[122]	https://pandas.pydata.org/ https://doi.org/10.5281/zenodo.10045529
PyTorch	The Linux Foundation; Paszke et al.[123]	https://pytorch.org/
scikit-learn	scikit-learn developers; Pedregosa et al.[124]	https://scikit-learn.org/stable/index.html
SciPy	Virtanen et al.[125]	https://scipy.org/
Spynal	Scott L. Brincat & John Tauber	https://github.com/sbrincat/spynal
Other		
PHD ULTRA 4400 computer controlled syringe pump	Harvard Apparatus	https://www.harvardapparatus.com/remote-infuse-withdraw-phd-ultra-4400-programmable-syringe-pumps.html
Vascular Access Port	Norfolk Access Technologies	https://norfolkaccess.com/our-products/clear-port/
8 x 8 iridium-oxide contact microelectrode arrays	Blackrock Microsystems	https://blackrockneurotech.com/products/utah-array/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Earl K. Miller (ekmiller@mit.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The electrophysiology data reported in this paper will be shared by the lead contact upon request.
- All figure data have been deposited at https://doi.org/10.5281/zenodo/11167882 and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code related to the analysis of electrophysiology data, simulation, and figure generation has been deposited at https://github.com/adamjeisen/ChaoticConsciousness and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code related to the stability estimation algorithm DeLASE has been deposited at https://github.com/adamjeisen/DeLASE and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

EXPERIMENTAL MODEL DETAILS

Multi-electrode neural activity was recorded from two rhesus macaques (Macaca mulatta), denoted NHP 1 and NHP 2 (standing for non-human primate). NHP 1 was female, aged 8 years, ~6.6kg at the time of the study and NHP 2 was male, aged 14 years, ~13.0-kg. Results were very similar between the male animal and female animal. Animals were pair-housed on 12 hr day/night cycles and maintained in a temperature-controlled environment (80 °F). The animals were not involved in previous procedures. Monkeys 1 and 2 were surgically implanted with a subcutaneous vascular access port (Model CP-6, Norfolk Access Technologies, Skokie, IL) at the cervicothoracic junction of the neck with the catheter tip reaching the termination of the superior vena cava via the external jugular vein. All procedures followed the guidelines of the Massachusetts Institute of Technology Committee on Animal Care and the National Institutes of Health.

The NHPs were implanted with 8 × 8 chronic Utah arrays, yielding 64 channels per multi-electrode array ('Utah arrays', MultiPort: 1.0 mm shank length, 400 mm spacing, Blackrock Microsystems, Salt Lake City, UT). The distance between electrodes ensured that we were not sampling spiking from the same neurons more than once. Electrodes were placed in four areas: ventrolateral prefrontal cortex (PFC), frontal eye fields (FEF), posterior parietal cortex (PPC) and auditory cortex (STG).[2] In the present analysis, we use LFPs from each area. During Utah array recordings, areas FEF and PFC were ground and referenced to a common subdural site. Areas STG and PPC also shared a common subdural ground/reference channel. The LFPs were recorded at 30 kHz and filtered online via a lowpass 250 Hz software filter. The LFPs were then subsequently downsampled to 1 kHz. To ensure synchronization of signals recorded on the multiple data acquisition systems, we simultaneously recorded a synchronization test signal with locally unique temporal structure on one auxiliary analog channel of each system. Throughout the entire recording session, we regularly measured offline the relative timing of this test signal between each of the system's recorded datafiles. Measured timing offsets between datafiles were rectified by appropriately shifting event code timestamps, and in addition by linearly interpolating analog signals to a common time base.

Note that for the stability analyses, since our method models population dynamics, all usable electrodes in each area were included in the analysis, and no averaging was performed. Each anesthetic infusion is referred to as a "session". By "session", we mean a continuous recording, from wakefulness through anesthetic infusion to recovery, in one animal. There are 21 anesthetic sessions across two non-human primate subjects (NHPs): 10 sessions with NHP 1, and 11 sessions with NHP 2. Sessions were always separated by at least 2 days, and were thus treated as independent samples. In each session, the NHP performed a non-demanding delayed saccade task pre-anesthesia. Then, following this task, the NHP experienced a passive airpuff/tone classical conditioning task. Specifically, on one third of trials the NHP was delivered a 500 ms ringing tone, which after a 500 ms blank delay, is followed by an airpuff (lasting about 10 ms) to the eye/face designed to elicit a blink response (notably, the airpuff also emitted a sound). In another third of trials there was an isolated airpuff with no tone, and on the last third of trials there was a distinct tone played that indicates that no airpuff will follow. After around 15-20 minutes of airpuff/tone classical conditioning the NHP was infused with propofol anesthesia for 60 minutes. 30 minutes at a higher loading dose (0.58 mg/kg/min for NHP 1 and 0.285 mg/kg/min for NHP 2) and 30 minutes at a lower maintenance dose (0.32 mg/kg/min for NHP 1 and approx. 0.075 mg/kg/min for NHP 2). The infusion was then stopped and the NHP gradually returned to a normal awake state. propofol was intravenously infused via a computer-controlled syringe pump (PHD ULTRA 4400, Harvard Apparatus, Holliston, MA).

All data analysis and simulation was done in Python using original code in addition to, Hydra, Matplotlib,[121] Pandas,[122] PyTorch,[123] scikit-learn,[124] scipy,[125] and spynal.

METHOD DETAILS

LFP Preprocessing

Recordings for each session were quite stable. Utah arrays have been shown to be stable on the order of days.[126] But to further ensure instability across sessions did not affect results, we first removed the mean across the entire session from each LFP. Then, to handle noise sources, we removed line noise via temporally windowed sinusoid fits at 60 Hz and all harmonics, as well as empirically found line noise frequencies: 107.35, 214.7, 190.2, 196.8, 393.6. We then low pass filtered each LFP at 300 Hz (3rd order bidirectional Butterworth). It was not necessary to consider eye movement and muscle artifacts in the preprocessing as these are considered to be signals that are not cerebral in origin,[127,128] and are more commonly considered in electroencephalography (EEG) and magnetoencephalography (MEG) studies due to their non-invasive nature.[129,130] The invasive intracortical electrophysiology signals used in this study are fortunately not susceptible to the same signal contamination risks as EEG and MEG. Furthermore, electrodes were not averaged at any point in the stability analysis pipeline. This is because our method models population dynamics and takes into account how the overall population representation transforms over time. That information is lost when electrodes are averaged.

Dynamics and stability estimation approach

Here we outline the DeLASE (Delayed Linear Analysis for Stability Estimation) approach to stability estimation.

Eigen-time-delay coordinate dynamics (HAVOK)

We consider observed data consisting of Tobservations of N dimensions (i.e. a matrix in $\mathbb{R}^{T\times N}$). Given a sampling interval of Δt, this corresponds to a window length of $T\Delta t$. Following HAVOK (Hankel Alternative View of Koopman), we construct a delay embedding, i.e. a matrix of the form

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

$$\mathbf{H} = \begin{bmatrix} \mathbf{x}_p & \mathbf{x}_{p+1} & \cdots & \mathbf{x}_T \\ \mathbf{x}_{p-1} & \mathbf{x}_p & \cdots & \mathbf{x}_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{T-p+1} \end{bmatrix}$$

Where $\mathbf{x}_t \in \mathbb{R}^N$ are the state observations at time t (e.g. channel activities), and p is the number of lags in the delay embedding matrix. Thus $\mathbf{H} \in \mathbb{R}^{Np \times (T-p+1)}$. We can now perform SVD on $\mathbf{H}$ to obtain $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}$ are the delay embedding's spatial modes, $\mathbf{\Sigma}$ are the singular values, and $\mathbf{V}$ are the eigen-time-delay coordinates (temporal modes). We select the r coordinates with largest singular values to obtain a reduced rank matrix $\mathbf{V}_r \in \mathbb{R}^{T \times r}$ of temporal modes. The benefits of this SVD step are twofold. First, it avoids a somewhat ill-posed regression, as most of the terms in each column of the delay embedding matrix $\mathbf{H}$ are identical to the previous column (but shifted). Second, it also removes any extraneous information present in the delay embedding due to correlation dimensions and noise. The computation of eigen-time-delay coordinates is equivalent to performing PCA whitening on the delay embedding matrix.[89] PCA whitening projects a data matrix into a space in which coordinates are decorrelated and have equal variance. This is explored later in the STAR Methods in the section "responses in eigen-time-delay coordinates". We then compute the dynamics matrix $\mathbf{A_V} \in \mathbb{R}^{r \times r}$ as the matrix that solves the following least squares regression problem:

$$\mathbf{A_V} = \mathrm{argmin}_{\mathbf{A}} \|\mathbf{V}_r^{+\mathrm{T}} - \mathbf{A}\mathbf{V}_r^{-\mathrm{T}}\|_2$$

Where $\mathbf{V}_r^- \in \mathbb{R}^{(T-p) \times r}$ is the matrix consisting of the first r eigen-time-delay coordinates at times 1 through $T - p$ and $\mathbf{V}_r^+ \in \mathbb{R}^{(T-p) \times r}$ is the matrix consisting of the first r eigen-time-delay coordinates at times 2 through $T - p + 1$. We can therefore construct a discrete model of the delay-embedded neural dynamical system as

$$\mathbf{h}_{t+1} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{A_V} \mathbf{\Sigma}_r^{\dagger} \mathbf{U}_r^{\mathrm{T}} \mathbf{h_t}$$

Where $\mathbf{h_t}$ is the column of the delay embedding matrix $\mathbf{H}$ corresponding to maximal time t (that is, including times $t - p + 1$ through t, $\mathbf{U}_r \in \mathbb{R}^{Np \times r}$ is the matrix corresponding to the first r columns of $\mathbf{U}$, $\mathbf{\Sigma}_r$ is the matrix corresponding to the first r singular values, and $\mathbf{\Sigma}^{\dagger}$ is the pseudoinverse of $\mathbf{\Sigma}$.

Estimating stability from delay dynamics

We now let $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{A_V} \mathbf{\Sigma}_r^{\dagger} \mathbf{U}_r^{\mathrm{T}}$ so $\mathbf{A}$ is the linear representation of the dynamics in the delay-embedded neural space. We now note that the first N rows of the matrix $\mathbf{A}$ describe the dependence of $\mathbf{x_t}$ on the trajectory history:

$$\mathbf{x}_t = \sum_{k=1}^{p} \mathbf{A}_k \mathbf{x}_{t-1}$$

Where $\mathbf{A}_k$ is the $N \times N$ matrix corresponding to columns kN through $(k+1)N$ of the first N rows of $\mathbf{A}$. We have thus arrived at a representation of the neural dynamics in the form of a (discrete) linear delay differential equation. To extract the stability from this representation of the dynamics, we turned to the field of delay differential equations. We first convert to continuous time (for purposes of interpretation) setting $\tilde{\mathbf{A}}_k = \frac{\mathbf{A}_k}{\Delta t}$ for $k > 1$ and $\tilde{\mathbf{A}}_k = \frac{\mathbf{A}_k - \mathbf{I}_N}{\Delta t}$ for $k = 1$ (this latter term is an approximation of the instantaneous term, and so it includes an identity matrix). This yields the delay differential equation $\dot{\mathbf{x}}(t) = \sum_{k=1}^{p} \tilde{\mathbf{A}}_k \mathbf{x}(t - k\Delta t)$. The stability of this delay differential equation is determined by the roots λ of its corresponding characteristic equation[92]:

$$\det\left(\lambda \mathbf{I}_N - \sum_{k=1}^{p} \tilde{\mathbf{A}}_k e^{-\lambda k \Delta t}\right) = 0$$

The intuition for why this equation is helpful in capturing the stability present in the system is because of the $e^{-\lambda k \Delta t}$ term. This term discounts the impact of the matrix $\tilde{\mathbf{A}}_k$ based on how far it is from the current state (Figure 2E).

This characteristic equation has infinitely-many solutions.[131] In fact, even for a simple delay differential equation with a single delay, the characteristic equation has infinite solutions. Intuitively, this is because in a continuous time setting there are infinite points in between the delay time and the current time - thus the system is infinite dimensional and must be specified with an infinite dimensional initial condition. Mathematically, the infinitely-many solutions are a result of a corollary to The Great Picard Theorem, namely that an entire non-polynomial function assumes every complex number infinitely many times (with one possible exception).[132] Analytic investigation of the stability of even single delay equations can necessitate quite advanced mathematical machinery.[133] There exist numerous methodologies to numerically approximate a finite portion of the roots of a given delay differential equation.[92,93,134,135] These approaches typically discretize the delay period and are guaranteed to converge to the true characteristic roots as the discretization of the delay period becomes finer. For our analysis, we use the TRACE-DDE algorithm, which estimates the roots of the characteristic equation by constructing a discrete approximation of the infinitesimal generator.[93] We choose the number of discretization points in the delay period to be equal to the number of lags chosen for the delay embedding, with discretization points spaced out by Δt. This generates $N(p+1)$ characteristic roots, of which for the present analyses we evaluate the top 10% (i.e. the 10% with greatest real part).

The characteristic roots are complex valued numbers. The real part of the root determines the rate at which perturbations to the system along a particular direction will decay. The complex part of the root determines the frequency at which such perturbations will

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

decay. In a strictly linear delay differential equation, the root with largest real part determines the stability, as it determines the overall slowest rate at which perturbations will decay (if the real part is negative) or the fastest rate at which they will explode (if the real part is positive). For our analysis, since we are approximating complex nonlinear systems with linear delay differential equations, we look at the upper portion of the distribution of characteristic roots extracted to characterize stability. Further, while the convention is to report the real parts of characteristic roots as inverse timescales, in our analysis we represent these values both with inverse timescales and standard timescales where appropriate.

To handle instabilities introduced into the dynamics by sensory stimuli, inputs that are not accounted for, noise, and other artifacts, we filter the extracted characteristic roots in two ways. First, we enable filtering out all characteristic roots with a frequency component above a certain value. For our analysis, we filtered out characteristic roots with a frequency component greater than 500 Hz, since this is the maximum possible frequency in data sampled at 1 kHz. Next, we enable filtering out all unstable roots with a frequency component above a certain value. For our analysis, we filtered out unstable roots with a frequency component greater than 125 Hz. The intuition for this is that rapid, sharp changes due to sensory inputs and other artifacts will destabilize the dynamics with a high frequency component, due to the sharpness of the change. We also note that when visualizing the characteristic roots as proper timescales in Figure 4B, we considered only the characteristic roots with negative real part. This was because the timescales associated with roots with negative real part are timescales of stable decay – in contrast to timescales associated with roots with positive real part, which are timescales of unstable explosion. Nearly all computed characteristic roots had a negative real part (see Figures 4C and 4E), and all statistics were performed on inverse timescales, which included any roots with positive real part.

Picking hyperparameters

To select (1) the minimum number of delay embedding coordinates and (2) the rank of the eigen-time-delay coordinates used to estimate the dynamics, a grid search was performed for each session (sample is shown in Figure S5A). We use fixed window sizes for our analysis (10 seconds for simulated data and 15 seconds for neural data). These window sizes ensure a long enough time window to capture a wide range of dynamic motifs, but a short enough time window to preserve computational tractability. For the grid search on neural data, models were fit on 12 windows from each session, and parameters were chosen on a per-session and per-area basis. Windows were chosen to ensure that all sections of the session were included (i.e. awake, anesthetic induction, anesthetic unconsciousness, and recovery), with 4 windows being randomly selected from each section in each session. The metric used to assess model quality was Akaike Information Criterion (AIC). AIC quantifies the balance between high quality model prediction (as measured by one-step prediction error on test data) with the number of model parameters. Models are penalized for having larger numbers of parameters, such that the optimal model of minimal complexity can be obtained. While larger models tend to perform better, after a certain point the added parameters are not benefitting the prediction quality sufficiently to justify their inclusion (Figure S5A, lower right corner of the grid). Because we compare stability between sections of each session, the hyperparameters minimizing AIC within a given session (and area) are chosen. The time interval between delays was held fixed at 1 time step (1 ms) for this analysis. Chosen hyperparameters varied across sessions but were typically around 750 delay coordinates (12 delays) with a rank of 750 for individual areas (usually ~64 channels per area), and approximately 1000 delay coordinates (4 delays) with a rank of 900 for all areas considered together (usually ~250 channels). All tested hyperparameter combinations preserve the core results of our analysis (see Figure S5B).

Implementation of VAR

We implemented VAR(1) (1st-order vector autoregression) in the following way. We again consider observed data consisting of T observations from N dimensions (i.e. a matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$). Given a sampling interval of Δt, this corresponds to a window length of $T\Delta t$. We then find the matrix $\mathbf{A}_{VAR} \in \mathbb{R}^{N \times N}$ that solves the following least squares regression problem:

$$\mathbf{A}_{VAR} = \text{argmin}_{\mathbf{A}} \|\mathbf{X}^{+\top} - \mathbf{A}\mathbf{X}^{-\top}\|_2$$

where $\mathbf{X}^{-}$ is the $(T-1) \times N$ matrix consisting of the N-dimensional state observations at times 1 through $T-1$ and $\mathbf{X}^{+}$ is the $(T-1) \times N$matrix consisting of the N-dimensional state observations at times 2 through T.

To estimate stability, we compute the eigenvalues of $\mathbf{A}_{VAR}$. To convert to a continuous time representation, for each eigenvalue λ_i we convert to the continuous time instability measure $\widehat{\lambda}_i = \frac{log(|\lambda_i|)}{\Delta t}$ as in previous work.[33] To estimate stability, we considered the largest 10% of such instability measures extracted from VAR, mirroring our approach for DeLASE.

Estimating stability in simulated systems

To validate the DeLASE method, and to generate examples of dynamical systems to compare to neural data, we simulated sample systems. The systems included simple linear dynamics as well as randomly connected RNNs. Since we were interested in systems that are stable, we often simulated these systems in the stable regime. In this regime, however, these systems converge to fixed points. Thus, to avoid this trivial behavior, we simulated these systems with stochasticity injected into their dynamics. In this approach, the systems were simulated with a small amount of process noise, effectively perturbing the system a small amount at each time step. We simulated stochastic dynamical systems using the Euler-Maruyama method. Specifically, we simulated the systems using the update:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{f}(\mathbf{x}_t)\mathrm{d}t + \sigma \mathbf{dW}_t$$

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

where $\mathbf{x}_t$ is the n-dimensional state of the system at time t, $\mathbf{f}$ is the system dynamics (i.e., $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$), d$t$ is the time step of simulation, σ is a small scale parameter on the noise, and $\mathbf{dW}_t$ is the n-dimensional process noise at time t, with each dimension independently sampled from a normal distribution with mean 0 and standard deviation $\sqrt{dt}$.

Linear dynamics

We simulated n-dimensional linear systems of the base form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is the $n \times n$ dynamics matrix. To set the maximal real part of the eigenvalues of this matrix, we first sample each matrix element from a normal distribution with mean 0 and standard deviation $\frac{1}{\sqrt{n}}$. Then we perform the update

$$\mathbf{A} = \mathbf{A} - \lambda_{max}\mathbf{I}_n + \lambda_{new}\mathbf{I}_n$$

Where λ_{max} is the original maximum real part of the matrix eigenvalues and λ_{new} is what we would like to update it to. We simulated the systems in the stochastic framework with $\sigma = 1$. For the simulations used to generate Figure 3A, we simulated stochastic linear dynamics with maximal eigenvalues from-1 to -0.1 (inclusive) stepping by 0.1. We simulated each system 20 times. Each system had 100 dimensions and was simulated for 20,000 time steps. The time step used was 0.002ms. A transient of 2000 time steps was dropped from the simulated systems, and models were fit to a randomly chosen partial observation of 10 dimensions of the system over 10,000 time steps. A hyperparameter grid search was executed to choose the number of delays and the rank of the eigen-time-delay coordinates. Specifically, minimum delay embedding matrix sizes of 10, 20, 50, 100, 200, 300, 500, 750 and 1000 were tested, along with ranks of 3, 5, 10, 25, 50, 75, 100, 125, 150, 200, all ranks from 200 to 800 stepping by 50, and also ranks of 900 and 1000. For each of the 20 runs, the best performing parameter combinations across all linear systems were chosen (typically a delay embedding matrix size of 10 with a rank of 10) and the stability was estimated based on these models.

Randomly connected RNN dynamics

We simulated randomly connected RNNs according to Sompolinksy et al.[97] Specifically, we simulated the dynamics:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) = \frac{\mathbf{1}}{\tau}(-\mathbf{x} + g\mathbf{W}\tanh\mathbf{x})$$

Where $\mathbf{x}$ is the n-dimensional state of the system (effectively synaptic activity in the RNN), τ is the time constant of the dynamics, typically set to 100 ms (with a time step of simulation of 10 ms), $\mathbf{W}$ is the $n \times n$ weight matrix with each element drawn from a normal distribution of mean 0 standard deviation $\frac{1}{\sqrt{n}}$, and g is a gain parameter on the synaptic weights. It has been shown that as n increases, and as the gain is increased above 1, the network enters into increasingly chaotic regimes.[97]

For the simulations presented in Figure 3B, we used a network size of 1024 neurons, and chose values for the gain parameter from 0.8 to 1.4 (inclusive), specifically: 0.8, 0.85, 0.9, 0.925, 0.95, 0.975, 1, 1.025, 1.05, 1.075, 1.1, 1.125, 1.15, 1.175, 1.2, 1.25, 1.3, 1.35, and 1.4. We used $\sigma = 0.05$. We sampled the matrix $\mathbf{W}$ 10 times, then for each sampled matrix simulated the dynamics with all possible values of the gain. The systems were simulated for 20,000 time steps. For stability analysis, a transient of 2000 time steps was dropped, and the models were fit to 10,000 time steps of data. We randomly observed 10 dimensions of the 1024-dimensional system to use for estimating stability. A hyperparameter grid search was executed to choose the number of delays and the rank of the dynamics matrix. Specifically, minimum delay embedding matrix sizes of 10, 20, 30, 40, 50, 100, 150, 200, 250, 300, 350, and 400 were tested with ranks of 2, 3, 5, 10, 30, 40, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, and 400. For each of the 10 runs, the best performing parameter combinations across all trajectories were chosen (typically a delay embedding matrix size of 50 with a rank of 40, though there was some variation) and the stability was estimated based on these models.

To compute the Lyapunov exponents in the simulated networks, we applied a QR-based algorithm[95] to the Jacobian of the discrete dynamics, computed as

$$\mathbf{J} = \mathbf{I}_\mathbf{n} + \frac{dt}{\tau}\left(-\mathbf{I}_\mathbf{n} + g\mathbf{W}\mathbf{diag}\left(\mathbf{1} - \mathbf{tanh}^\mathbf{2}\,\mathbf{x}\right)\right)$$

The exponents extracted from the QR algorithm are then converted into continuous time by dividing by the time step of simulation, dt.

Linear dynamics with inputs

One-dimensional linear sine wave filters

For Figure 5, we simulated a simple one-dimensional linear system with input according to the dynamics $\dot{x} = -\lambda x + u(t)$ where λ controls the instability of the filter. For the simple filter on a sine wave (Figure 5B), we found the analytical solution for the (deterministic) differential equation with $u(t) = \sin 2\pi\omega t$. The solution is $x(t) = \frac{1}{\lambda^2+1}(\lambda \sin 2\pi\omega t - \cos 2\pi\omega t)$. We computed this solution for values of λ from 0 to 4.5 (inclusive), stepping by 0.5, and $\omega = 1$ (Figure 5B). To simulate the dynamics in the presence of noise (Figure 5C), we simulated simple one-dimensional stochastic dynamics:

$$x_{t+1} = x_t + (-\lambda x_t + u(t))dt + \sigma dW_t$$

with λ set to 10 and 0.01, and σ set to 0.05. We used a time step of 1 ms to match the sampling rate of the neural data, and a simulation time of 3 seconds. 1 second into the simulation, we perturbed the above equation with a negative 2 Hz sine wave for 500 ms ($u(t) = \sin 4\pi t$). This matches the duration of the tone presentations in the experimental setup. We then provided no inputs for 500 ms

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

before providing another negative 2 Hz sine wave input for 500 ms (this matches the spacing of tones and puffs in the experimental setup).

Two-dimensional linear system input responses

In the two-dimensional case, we simulated the two-dimensional stochastic dynamics given by:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + (\mathbf{A}\mathbf{x}_t + u(t)\mathbf{I}_2)dt + \sigma dW_t$$

To generate the dynamics $\mathbf{A}$, we took two parameters, a and b, as well as an orthogonal matrix $\mathbf{Q}$, and set

$$\mathbf{A} = \mathbf{Q}\begin{bmatrix} a & -b \\ b & a \end{bmatrix}\mathbf{Q}^{-1}$$

Thus $\mathbf{A}$ has eigenvalues $a \pm bi$ and a controls the stability of the dynamics. For Figure 6B, we simulated the dynamics with both $a = -6$ and $a = -1$. We set $b = 2\pi$, and picked $\mathbf{Q}$ randomly. We used a time step of 0.25 ms and a simulation time of 7 seconds and set $\sigma = 0.005$. The first input was a constant function $u(t) = 0.6$ lasting 500 ms and starting at 2 seconds. The second input was a constant function $u(t) = 1$ lasting 150 ms and starting at 3 seconds. The structure of the inputs was chosen to mirror the tones and puffs in the experimental setup.

Neural trajectories of stimulus responses

To compute the neural trajectories depicted in Figures 5D, 5E, and 6C–6H, we first compiled population responses from every session in each NHP for the trial type of interest (a 500 ms tone, followed by a 500 ms delay and then an air puff). To ensure enough resolution around the stimuli, we collected LFPs from 2 seconds before the first stimulus onset until 5 seconds after. For each NHP, this yielded a $K \times T \times N$ matrix, where K is the number of trials (aggregated across sessions), T is the number of time-points from the trial (as mentioned, we consider 7 seconds surrounding the stimulus onset), and N is the number of electrodes.

Average LFP responses

To compute the one-dimensional average LFP trial response (Figures 5D and 5E), for each NHP, we averaged the trial responses over trials, and then subsequently over electrodes.

Responses in eigen-time-delay coordinates

To compute the (scaled) eigen-time-delay coordinates of the average stimulus response (Figures 6C–6H), we first delay embedded each trial individually. Specifically, for a specific trial response, given that the neural population state at time t is $\mathbf{x}_t$, and given the delay embedding parameters (the delay interval τ and delays), we computed the $Np \times (T - (p-1)\tau)$ matrix given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{x}_{1+(p-1)\tau} & \mathbf{x}_{2+(p-1)\tau} & \cdots & \mathbf{x}_T \\ \mathbf{x}_{1+(p-2)\tau} & \mathbf{x}_{2+(p-1)\tau} & \cdots & \mathbf{x}_{T-\tau} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{T-(p-1)\tau} \end{bmatrix}$$

After performing a delay embedding for every trial, for each NHP we end up with a $K \times Np \times (T - (p-1)\tau)$ matrix. We then average across trials to arrive at a $Np \times (T - (p-1)\tau)$ matrix of the average delay embedding. Finally, we perform PCA on (the transpose of) this matrix to reduce the number of dimensions from Np to 2. Since each dimension of the delay embedding matrix contains observations from multiple time-points, we consider the "effective" time of the delay embedding coordinate to be the latest time-point in the dimension. That is, the effective time of the first column of $\mathbf{H}$ above is $1 + (p-1)\tau$. These are the times used for plotting in Figures 6C–6H.

Furthermore, we note again here that this process equates to computing scaled versions of the eigen-time-delay coordinates used by the HAVOK models. To see why this is the case, consider a delay-embedded matrix $\mathbf{X} \in \mathbb{R}^{N' \times T'}$ where N' is the number of coordinates in the delay embedding and T' is the number of delay embedded time-points. Thus, given the singular value decomposition $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, the first r eigen-time-delay coordinates used by HAVOK are the first r columns of $\mathbf{V}$. Note that by rearranging this equation we can show that the first r eigen-time-delay coordinates can be computed as $\Sigma_r^{-1}\mathbf{U}_r^T\mathbf{X} = \mathbf{V}_r^T$, and equivalently $\mathbf{V}_r = \mathbf{X}^T\mathbf{U}_r\Sigma_r^{-1}$ where the r in the subscript denotes the rank-r column truncation. Note here that $\mathbf{X}^T\mathbf{U}_r$ is the expression that computes the projection of the data onto the top r principal components in PCA (assuming the N' rows of $\mathbf{X}$ are mean-centered). Thus, in the eigen-time-delay coordinates, the diagonal matrix Σ_r^{-1} alters the PCA only in that it scales each component by the inverse of its corresponding singular value. This is known as PCA whitening and produces a scaled version of the data projected onto the principal components such that each component has unit variance.

Increasing inhibition in RNNs

To examine the effects of increasing inhibition (Figure 7), we again simulated randomly connected RNNs according to Sompolinksy et al.[97] following the dynamics

$$\tau\dot{\mathbf{x}} = -\mathbf{x} + g\mathbf{W}\tanh\mathbf{x}$$

where **x** is the n-dimensional state of the system (effectively synaptic activity in the RNN), τ is the time constant of the dynamics (10 ms, with a time step of simulation of 1 ms), **W** is the $n \times n$ weight matrix with each element drawn from a normal distribution of mean 0 standard deviation $\frac{1}{\sqrt{n}}$, and $g$is a gain parameter on the synaptic weights. We used a network size of 512 neurons. We simulated the dynamics with gains 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7. To scale up and down the inhibition in the network, we took all elements of **W** with negative weight, and scaled them by an inhibitory scaling factor, κ_I. For each gain, we simulated the network with $\kappa_I = 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4,$ and 1.5. We sampled the matrix **W**10 times, then for each sampled matrix simulated the dynamics with all possible pairings of g and κ_I. The systems were simulated for 10,000 time steps. We computed the Lyapunov exponents of the networks as described above in the section "randomly connected RNN dynamics".

Example dynamical systems

HAVOK models of the Van der Pol Oscillator

We simulated the Van der Pol Oscillator[91] with a time step of 20 ms and a simulation time of 400 seconds (generating 20,000 time steps) using the governing equations

$$\dot{x} = y$$

$$\dot{y} = \mu(1 - x^2)y - x$$

We set the parameter μ to 2. We chose an initial condition from a standard Gaussian distribution then dropped the first 5,000 time steps (100 seconds) as a transient. We then observed only the second (y) dimension of the dynamics. We fit a HAVOK model on 10,000 time steps (200 seconds) using 500 delays. Predictions were generated on the remaining 5,000 time steps by autonomously running the dynamics - that is, the initial state is fed through the model to get the first prediction, then that prediction is fed back through the model, and so forth. The results of this simulation are presented in Figure 2D.

Mass-spring simulation

For the cartoon shown in Figure 5A, a mass-spring system was simulated with governing dynamics $m\ddot{x} + R\dot{x} + kx = I$ where x is the vertical position of the mass, m, k, R are parameters governing the dynamics and I is the input. For our simulations, we set $m = 10$, $k = 1, R = 10$. We simulated using scipy's *solve_ivp* function for 1,000 time steps with a time step of 10 ms.[131] From time step 100 to 120, we perturbed the system with a constant input of value $I = 10$. The results of this simulation are in Figure 5A (right).

Pharmacokinetics analysis

Estimating integrated propofol dosage

To estimate the integrated (i.e., accumulated) propofol dosage up to a particular moment in the session, we began by estimating the context-sensitive half time (CSHT) of propofol. CSHTs describe the time it takes for the concentration of anesthetic in the system to halve, as a function of the duration of infusion (i.e., the "context"). Detailed modeling has estimated that the CSHT of propofol is around 3 minutes for a very short infusion and around 8 minutes after 60 minutes of infusion.[136] We then linearly interpolated between these two values for infusion times in between 0 minutes and 60 minutes, and assumed the CSHT stays constant at 8 minutes after the 60 minute infusion ends. Given the CSHT, we can estimate the instantaneous time constant of decay of propofol by assuming that the quantity present in the system undergoes exponential decay in the absence of any input. We thus obtained, using the definition of the CSHT, the expression

$$0.5x_0 = e^{-\frac{\mathrm{CSHT}(t)}{\tau}}x_0$$

where x_0 is the initial quantity present in the system, τ is the instantaneous time constant of decay in minutes, and $\mathrm{CSHT}(t)$ is the CSHT after t minutes of infusion. We can solve this equation to obtain that $\tau = \tau(t) = -\frac{\mathrm{CSHT}(t)}{\log(0.5)}$ minutes. Now, we integrate, over the course of a full session, a first-order differential equation for the quantity of propofol present in the system:

$$\dot{x} = \alpha - \frac{1}{\tau(t)}x$$

Where x is the propofol quantity in mg/kg, $\dot{x}$ is the rate of change of propofol quantity in mg/kg/min, and α is the infusion rate in mg/kg/min (see Experimental Model Details). We solve this equation using scipy's *solve_ivp* function. Given the complexity of propofol pharmacokinetics,[137] we stress that these estimates were intended only to be approximations for the purposes of investigating the relationship between propofol dosage and instability.

Predicting instability from propofol dose

Using the estimated integrated propofol dosage, we attempted to predict the estimated instability values. For each window in each session, we computed both the estimated integrated propofol dose up to that time, as well as the mean instability in that window. Then, for each NHP, we fitted a linear-log regression to predict instability from the estimated integrated propofol dose. The results of this analysis are in Figure S2D.

Please cite this article in press as: Eisen et al., Propofol anesthesia destabilizes neural dynamics across cortex, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.06.011

QUANTIFICATION AND STATISTICAL ANALYSIS

For all data plots, "*" denotes a p-value less than 0.05, "**" denotes a p-value less than 0.01, and "***" denotes a p-value less than 0.001.

Prediction quality of dynamical models

To quantify the prediction quality of the dynamical models considered in the manuscript (Figure 3; section "delayed linear dynamical systems models capture neural dynamics"), we first fit each model for every window of every session. Then, the performance metrics (MSE and AIC) were averaged across windows within a single session. To compute the statistics, each session was an independent sample. Since, for each session, we were comparing between two models, we used the one-sided Wilcoxon signed-rank test. Since each session was an independent sample, the sample size was $n = 10$. For Figure S1, we split the analysis by NHP, yielding sample sizes of:

- NHP 1: $n = 10$
- NHP 2: $n = 11$

Changes in stability in neural dynamics

Statistical tests referenced in the section "propofol anesthesia induces unstable cortical dynamics" were conducted as follows. In each NHP, the real parts (i.e. inverse timescales of response) of the top 10% of characteristic roots from every window in every session were grouped according to the section (i.e. awake, unconscious, recovery, early recovery, late recovery, loading dose, and maintenance dose) of the session. The mean of these inverse timescales was then taken to yield one value per section, per session.

To compute statistics, the sections were defined as follows. Awake windows were collected from the 15 minutes preceding the start of the anesthetic infusion. Induction windows were collected from the start of propofol infusion until 15 minutes following the start of propofol infusion. Unconscious windows were collected from 15 minutes following the start of propofol infusion until the end of propofol infusion (a period of 45 minutes). Recovery windows were collected from the end of propofol infusion until 15 minutes following the end of propofol infusion. Early recovery was collected from the end of propofol infusion until 8 minutes following the end of propofol infusion. Late recovery was assessed from 8 minutes following the end of propofol infusion until 15 minutes following the end of propofol infusion. Loading dose windows were collected from 15 minutes following the start of propofol infusion until 30 minutes following the start of propofol infusion (the final 15 minutes of the loading dose). Maintenance dose windows were collected from 45 minutes following the start of propofol infusion until the end of propofol infusion (the final 15 minutes of the maintenance dose). See Figure S6 for a visual depiction of the section definitions.

To test whether one section of the session was more unstable than another, one-sided Wilcoxon signed-rank tests were performed on mean instability values (top 10% of the real parts of characteristic roots) and averaged within each section for each session. Thus for each section, for every area, the sample sizes are:

- NHP 1: $n = 10$
- NHP 2: $n = 11$

The results of all comparisons between sections are reported in the main text in the section "propofol anesthesia destabilizes cortical neural dynamics". Most of these comparisons are also visually presented in Figure 4D.

Distances between stability distributions

To test whether the recovery distribution returned to awake levels of dynamic stability, we first constructed a test value that consisted of the awake-unconscious Cramér-von Mises criterion, subtracted from the awake-recovery Cramér-von Mises criterion. The Cramér-von Mises criterion was chosen as it constructs a non-parametric notion of distance between distributions that considers the whole distribution (as opposed to the widest gap, as is the case with Kolmogorov-Smirnov distance). To compute the Cramér-von Mises criterion, all characteristic roots from a given section (i.e., awake, unconscious, etc.) of a single session were included to ensure the maximum resolution. This approach yielded one test value per session. With this construction, if the test value is negative, it suggests that the recovery distribution is closer to the awake than unconscious state, indicating a return to awake stability levels. We then performed a one sample Wilcoxon signed-rank test on these differences to test if they were significantly below zero (i.e., if the test statistic was indeed negative). The sample sizes are the same as above. The results of this analysis are reported in the main text in the section "propofol anesthesia destabilizes cortical neural dynamics".

Changes to characteristic root frequencies

To analyze how the frequency (imaginary) components of the top 10% of characteristic roots changed in anesthesia, we first aggregated the top 10% of characteristic roots from all windows from all sessions in each NHP. We used the roots estimated from models fit to all areas considered together. Then, to test whether the proportion of roots in each band changed, we computed the proportion of characteristic roots that fell into that band during both the awake state and the unconscious state. This is equivalent to calculating

the probability mass function for each band. To test whether the mass was significantly different between the awake and unconscious states, for each band, we performed a two-sided Wilcoxon signed-rank test on the computed probability mass for that frequency band. Sample sizes are the same as above. The results of this analysis are in Figure S2E.

Destabilization in simulated networks

As mentioned above, for each sampling of the weight matrix **W**, we simulated every possible pairing of the synaptic gain g and inhibitory scale κ_I. Thus, to test whether a given inhibitory scale κ_{I_1} induced more instability than another inhibitory scale κ_{I_2}, we compared the maximum Lyapunov exponents of the simulations from each gain and from each sampling of the weight matrix. Specifically, given a particular sample of **W** and value of g, we can consider the two networks with inhibitory scales κ_{I_1} and κ_{I_2} to be paired dependent samples. Thus, we can conduct a one-sided Wilcoxon signed-rank test using the 100 paired samples for (10 samples of **W** times 10 values of g) for each pair of inhibitory scales. The results of this analysis are in Figure S4.