HHS Public Access

Author manuscript

Nat Neurosci. Author manuscript; available in PMC 2018 June 04.

Published in final edited form as:

Nat Neurosci. 2018 January ; 21(1): 102–110. doi:10.1038/s41593-017-0028-6.

Flexible timing by temporal scaling of cortical responses

Jing Wang[*,1], **Devika Narain**[*,1,2], **Eghbal A. Hosseini**[2], and **Mehrdad Jazayeri**[1,2]

[1]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Abstract

Musicians can perform at different tempos, speakers can control the cadence of their speech, and children can flexibly vary their temporal expectations of events. To understand the neural basis of such flexibility, we recorded from the medial frontal cortex of nonhuman primates trained to produce different time intervals with different effectors. Neural responses were heterogeneous, nonlinear and complex, and exhibited a remarkable form of temporal invariance: firing rate profiles were temporally scaled to match the produced intervals. Recording from downstream neurons in the caudate and thalamic neurons projecting to the medial frontal cortex indicated that this phenomenon originates within cortical networks. Recurrent neural network models trained to perform the task revealed that temporal scaling emerges from nonlinearities in the network and degree of scaling is controlled by the strength of external input. These findings demonstrate a simple and general mechanism for conferring temporal flexibility upon sensorimotor and cognitive functions.

Introduction

Mental capacities such as anticipation, motor coordination, deliberation, and imagination lie at the heart of higher brain function. A fundamental feature of these capacities is that they are not tied to immediate sensory or motor events and unfold at different timescales. To support such temporal flexibility, the brain must control the dynamics of ongoing patterns of neural activity. An example of such flexible behavior is the control of self-initiated

Corresponding author statement: Correspondence may be directed to MJ : mjaz@mit.edu.

Present Addresses: JW, EH and MJ: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA USA

DN: Netherlands Institute for Neuroscience, Amsterdam, The Netherlands; Department of Neuroscience, Erasmus Medical Center, Rotterdam, The Netherlands

[*]Equal contribution statement: JW and DN contributed equally to this research

Author Contributions

J.W. was responsible for all aspects of experiments, analyses and developed the simplified model. D.N. was responsible for the development of the recurrent neural network model. E.A.H. helped with the data collection and analysis. M.J. was responsible for all aspects of the project. All authors helped with the interpretation of data and writing the paper.

Competing financial interests statement

The authors declare no conflicting interests.

movements. Humans can precisely control the timing of their movements and can make rapid adjustments based on instruction. However, the mechanisms that confer such flexibility are not well-understood.

We investigated the neural mechanisms underlying flexible temporal control. We developed a task in which monkeys were instructed to produce different time intervals using different effectors. While monkeys performed the task, we evaluated the causal function and signalling properties of neurons across three brain areas that have been strongly implicated in timing: 1) the dorsomedial frontal cortex (MFC), which has been implicated in the inhibition[1], initiation[2,3], and coordination[4–7] of movements, 2) the caudate nucleus downstream of MFC that is thought to play a major role in timing tasks[8–15], and 3) thalamic regions that project to MFC and causally influence self-initiated movements[16].

Neurons exhibited a diversity of complex response profiles that could not be reconciled with dominant models of timing[13], including the clock-accumulator models[17,18], the oscillation-based models[19], and the population clock models[20,21]. Instead, responses were unified under a general principle of temporal scaling that was evident at both individual and population levels. Specifically, when animals produced longer intervals, the population activity evolved along an invariant neural trajectory but at a slower speed. Remarkably, speed was adjusted on a trial-by-trial basis and in accordance with the instruction provided to the animal. Although these findings are at odds with classic models of timing, they corroborate observations of temporal scaling in other tasks and areas[8,22–25].

To investigate the mechanisms underlying such flexible speed control, we analyzed the dynamics of recurrent neural network models capable of using a graded input to produce different time intervals. Analysis of these models revealed a novel yet simple mechanism for flexible temporal scaling: degree of scaling was controlled by an external input acting upon the nonlinear activation function of individual neurons in a recurrent network.

Results

Behavior

On each trial, monkeys fixated a central spot with their hand resting on a button, and produced either a *Short* (800 ms) or a *Long* (1500 ms) interval using one of the effectors (*Eye* or *Hand*). The desired interval and effector changed on a trial-by-trial basis and was cued throughout the trial by the color and shape of the fixation point (Fig. 1a). Production intervals (*Tp*) were measured from a brief ‘Set’ flash to the time of movement initiation. Animals learned to flexibly switch between conditions (Fig. 1b) and produced accurate intervals whose variability increased for the *Long* condition compared to *Short* (Fig. 1c). This is consistent with Weber’s law and is a well-known property of timing behavior[26,27]. The Weber fraction was significantly larger for button presses compared to saccades (one-tailed paired sample t-test, for monkey A, n = 31, $t_{30} = 1.80$, $P = .041$, and for monkey D, n = 35, $t_{34} = 6.44$, $P < .001$).

Causal experiments and single-unit electrophysiology

Reversible inactivation of MFC (Fig. 2a) with muscimol, a $GABA_A$ agonist, significantly impaired performance for both *Long* and *Short* intervals (Fig. 2b). This was evident from a comparison of the distribution of within-session increases in the mean-squared error (MSE) after the muscimol injection, when compared to before (for statistics, see Table 1). The drop in performance was due to a combination of changes in both bias and standard deviation (Table 1). No significant impairment was measured after saline injection (Fig. 2b and Table 1). Furthermore, muscimol inactivation had no significant effect on reaction times during a memory saccade task (Table 1). Based on these results, we concluded that MFC played a causal role in the main motor timing task.

Temporal scaling of complex response profiles

To estimate each neuron's firing rate, we binned trials based on *Tp* and computed average spike counts after aligning trials to the time of the motor response (Fig. 2c). Across neurons, response profiles were highly heterogeneous and included linear, nonlinear, monotonic, non-monotonic and multi-modal activity profiles (Fig. 2d). We tested each neuron's activity profile against predictions of various models of motor timing using a cross-validation procedure (Fig. 2e). We considered three variants of the clock-accumulator model, one in which flexible timing was achieved by adjusting a threshold over a ramping process, one in which the clock was adjusted, and one in which both were adjusted. Since clock models can only accommodate neurons with linear ramping profiles[17,18,28–30], they failed to capture the nonlinear profiles exhibited by the majority of neurons in the population. Cross-validated polynomials fits of different degrees to response profiles substantiated that only 11% (47/416) of responses increased linearly; the rest were explained by higher order polynomials. This number increased by only 4% when we allowed the starting and terminating points of the linear ramp to vary by up to 200 ms.

We also tested two oscillation-based models of interval timing, in which the response time is determined by the collective phase of oscillators with different frequencies[19]. In one variant, a single sinusoid was fit to the response of each neuron, and in another, multiple sinusoids (up to 4) of different frequencies were used. These models were also unable to capture the diversity of MFC responses (Fig. 2e).

Finally, we tested MFC responses against a simple variant of the population-clock model[20,21] in which the collective activity of a population of neurons with unique and context-independent response profiles controls movement initiation time. Accordingly, we modeled each neuron by the best-fitting polynomial (cross-validated) that captured the activity across both the *Short* and *Long* contexts. This model performed better than the clock-accumulator and oscillation models. However, MFC data violated a key qualitative prediction of the population clock model: unlike the population clock model, vast majority of MFC responses differed for the *Short* and *Long* conditions from early on after the Set cue (Fig. 2d).

Our initial inspection indicated that response profiles were self-similar when stretched or compressed in accordance with the produced interval (Fig. 2c, d, and Supplementary Fig. 1).

This was true both for fluctuations of *Tp* within each temporal context (i.e., 800 ms or 1500 ms), and across the two temporal contexts. Consistent with this observation, a temporally-scaled polynomial function fitted to the data for different conditions clearly outperformed all other models in terms of explanatory power (Fig. 2e, one-way ANOVA, $F_{6, 2859} = 125.2$, $P < .001$).

Speed control across the population

We quantified the degree of scaling by a *scaling index* (SI) that was computed as a coefficient of determination (R^2) across temporally-scaled responses associated with different *Tp* bins. This analysis revealed a wide range of SI values across the population (Supplementary Fig. 1a).

Speed adjustment at the population level

When activity of a population of neurons is plotted in a coordinate system in which each axis represents the firing rate of one neuron, also known as the *state space*, the response dynamics of the population can be depicted as a high-dimensional *neural trajectory*. In this representation, perfect temporal scaling would result in perfectly overlapping neural trajectories evolving at different speeds. When we plotted MFC neural trajectories within the space spanned by the first three principal components (PCs) of neural activity, responses did not overlap perfectly indicating that MFC responses comprised a mixture of scaling and non-scaling signals (Fig. 3a), which was also evident from the distribution of SI values across individual neurons (Supplementary Fig. 1a).

We hypothesized that perfect scaling might be found within a subspace of the population activity; i.e., *scaling subspace* (Fig. 3b). As a first step, we examined the degree of scaling in the first few PCs. Using the same SI metric used for single neurons, we found that the first two PCs that explained nearly 40% of the variance (Fig. 3b, bottom) had a scaling index of 0.91 and 0.97, respectively (Fig. 3a, bottom). The third PC, however, did not exhibit temporal scaling and had a SI of 0.20. This provided initial evidence that certain high-variance dimensions in the state space exhibit strong scaling. However, scaling dimensions need not coincide with PCs, since PCs correspond to dimensions of maximum variance – not maximum scaling. To identify the scaling dimensions, we developed a novel dimensionality reduction technique that furnished a set of *scaling components* (SCs) that were ordered according to the degree of scaling in the data (see Methods).

The SI values for the first few SCs were relatively large indicating that the optimization process correctly identified the scaling dimensions (Fig. 3c, bottom, Supplementary Fig. 2). Because SCs were cross-validated, the scaling index for SCs of the test data did not follow a strictly decreasing order, although this was the case for the dataset used to determine the SCs (not shown). Responses projected onto the subspace spanned by the first three SCs traced nearly identical trajectories that evolved at different speeds (Fig. 3c, top), which is precisely what is expected in the scaling-subspace.

Next, we asked how much variance in the neural data can the scaling subspace account for. Ordered SCs explained less variance than the corresponding PCs suggesting that the scaling dimensions were not identical to PC dimensions (Fig. 3c). To better quantify the relationship

between scaling and variance explained, we performed two complementary analyses. First, we examined the relationship between SI and variance explained for each SC. This analysis provided an initial evidence that SCs with large SIs explained a relatively large percentage of variance (Fig. 3d). Second, we developed a procedure for quantifying the relationship between scaling and variance without relying on projections onto specific directions, such as PCs or SCs. We used a bootstrap procedure and quantified the relationship between variance explained and SI along 200 random projections in the state space. We then constructed a two-dimensional probability distribution of the relationship between variance explained and SI across those random projections (Fig. 3d, inset). This analysis verified that the dimensions with large degrees of scaling also explained a large portion of the variance.

To validate SI as a reliable metric for scaling, we quantified SI for surrogate data created from Gaussian processes. The surrogate data was constructed to statistically match MFC responses in terms of smoothness, starting/terminal firing rates, dimensionality, and the correlation between *Short* and *Long* activity profiles but was not constrained to exhibit temporal scaling (see Supplementary Note and Supplementary Fig. 3). The surrogate data, despite being matched to the statistics of MFC responses, had smaller SIs than those computed for MFC neurons (Fig. 3e). This verified that a significant portion of variance in MFC resides within a scaling subspace in which activity evolves along invariant trajectories at different speeds.

Finally, we quantified the relationship between speed in the scaling subspace and behavior. Using cross-validation, we derived the scaling subspace from a subset of shortest and longest trials, and asked whether the speed of neural trajectories of the remaining trials in that subspace could predict *Tp*. Results indicated that longer *Tp*s were associated with slower speeds (Fig. 3f and Supplementary Fig. 4), and the average speed was inversely proportional to *Tp* ($R^2 = 0.87$). These results suggest that the brain controls the speed of neural trajectories in order to flexibly produce different time intervals. Importantly, this speed control seemed to explain both behavioral variability within each temporal context, and flexible switching between the two contexts.

Speed control across cortico-basal ganglia circuits

Having established speed control in MFC as a potential mechanism for temporal flexibility, we asked whether this property was also present downstream of MFC in the basal ganglia. We focused on a region of the caudate that is thought to receive direct input from MFC[31,32] (Fig. 4a–b). First, we used reversible inactivation to verify the causal involvement of this region in the task (Fig. 4b, statistics: Table 1). Afterwards, we recorded from individual neurons (Fig. 4c) and analyzed their responses with respect to the temporal scaling property. Caudate responses, like those in MFC, were complex and heterogeneous, and had different profiles for *Short* and *Long* trials. At the level of single neurons, the degree of scaling in the caudate was similar to MFC (Supplementary Fig. 1). At the population level, analysis of PCs and SCs verified the presence of a scaling subspace in the caudate (Fig. 3e, Supplementary Fig. 5). Finally, the SI values of PCs as well as an unbiased analysis of responses across random projections in the state space indicated that dimensions with strong scaling explained a large part of variance in the data (Fig. 4d). These analyses verified that neural

signals in the caudate shared the same key properties with MFC and could contribute to subspace speed control.

In addition to receiving inputs from MFC, the basal ganglia also projects back to MFC through the thalamus. The presence of this anatomical substrate raises the possibility that MFC inherits temporal scaling from the basal ganglia via transthalamic projections. To test this possibility, we targeted a region of the thalamus where MFC-projecting thalamocortical neurons were identified antidromically (Fig. 4e; see Methods). Consistent with previous work[16], reversible inactivation strongly influenced animals' timing behavior (statistics: Table 1). However, several observations indicated that the function of thalamocortical signals was different from that of the caudate and MFC (Fig. 4g). First, SIs of single thalamic neurons ($n_{thalamus}$ = 846) were significantly smaller across the population compared to the other areas (n_{MFC} = 416 and n_{Cd} = 278, Mann-Whitney-Wilcoxon test, $W(1260)$ = 310,733, z = 7.89, and P < .001 for MFC and $W(1120)$ = 189,163, z = 6.98, P < .001 for the caudate, see Supplementary Fig. 1a). Second, scaling in the thalamus was significantly smaller than the surrogate data (one-tailed two sample t-test, n = 200, t_{398} = 35.2, P < .001 comparing to C+D+E+S surrogate model, Fig. 3e). Third, scaling was less prominent in the thalamus as indicated by the relationship between the magnitude of scaling and variance explained along random projections in the state space (Fig. 4h). Fourth, unlike the caudate and MFC, neural trajectories in the thalamus were not invariant in the space spanned by the first three SCs (Supplementary Fig. 5). This was also evident in the profile of the second PC, which systematically changed in average value as opposed toscaling. Together, these observations provide strong evidence that thalamic neurons exhibit significantly less scaling than the MFC neurons they project to. Since the output of the basal ganglia to cortex is routed through the thalamus, the weak scaling in thalamocortical neurons implicates that scaling may originate within MFC or in other cortical circuits projecting to MFC.

A model for flexible subspace speed control

Since the timescales of MFC response modulations were slower than the intrinsic time constants of single neurons, we assumed that the observed dynamics were the result of network-level interactions. Motivated by recent advances in understanding the dynamics of cortical population activity using network models[33–35], we used a recurrent neural network model to investigate the potential underlying mechanisms of speed control (Fig. 5). The model received a context input (Cue) whose magnitude specified the desired interval and a transient pulse (Set) that cued the start of the interval (Fig. 5a). The network was trained so that its output (a weighted linear sum of its units) had to breach a fixed threshold at that the desired time[36].

The network learned to generate the desired output function (Fig. 5d) and the activity of model neurons emulated the key features observed in MFC: response profiles of individual network units were heterogeneous, complex and temporally-scaled (Fig. 5b). Moreover, the speed of population dynamics directly determined the produced interval (Fig. 5c). These observations were robust regardless of whether the training objective was linear, nonlinear, scaling or non-scaling (Supplementary Fig. 6). The scaling behavior also persisted when the

Cue input was provided transiently (Supplementary Fig. 6). Motivated by the robustness and generality of these results, we reverse engineered the networks to investigate the underlying mechanisms of temporal scaling.[37]

Temporal scaling could be explained in terms of a pair of input-dependent stable fixed points, F_{init} and $F_{terminal}$. At the start of the trial, the Cue initialized the state of the network to an inital fixed point, F_{init}. Activation of the Set pulse drove the system away from F_{init} allowing the system to evolve toward $F_{terminal}$ with a speed that was determined by the magnitude of the Cue input (Fig. 5c,e). Within the network, the input and the recurrent dynamics played complementary roles (Fig. 5c). The input specified the position of the initial and terminal fixed points along a direction, which we refer to as the *input subspace*. Recurrent dynamics on the other hand, established a *recurrent subspace*, which determined the neural trajectory between these fixed points. These two subspaces emerged from different components of the network. The input subspace was governed by the direction specified by the input weights. In contrast, the recurrent subspace emerged from the constraints imposed by the recurrent weights. The two subspaces also differed in terms of their relationship to the scaling phenomenon. Within the input subspace, different intervals were associated with a change in the level of activity but did not exhibit scaling. This change in level controlled the speed by setting the position of the neural state along the axis of the input subspace. The recurrent space, on the other hand, did not control the speed but was responsible for the emergence of invariant trajectories and temporal scaling.

The division of labour between these subspaces provides a remarkable and unsuspected explanation of why scaling and non-scaling signals might coexist within the same network. Non-scaling signals reflect the input that sets the speed, and scaling signals correspond to the evolution of activity with the desired speed. This organization predicts that MFC neurons with weak temporal scaling are likely recipients of relatively strong context-dependent input, possibly derived from signals in upstream thalamic neurons (Fig. 4g), and neurons with strong temporal scaling are more directly engaged in recurrent interactions. Finally, the model-based distinction between these two subspaces provides a theoretical basis for analyzing MFC responses within a scaling subspace, which corresponds to the recurrent subspace in the model.

Importantly, the model allows us to infer that within the non-scaling input subspace, production times should be correlated with the average level – not speed – of neural activity. To test this novel prediction, we investigated whether *Tp* could be predicted by the non-scaling component of MFC activity. We inferred the least-scaling direction from our scaling component analysis. SCs specified an orthonormal basis whose axes were ordered according to the level of scaling (Supplementary Fig. 7). Therefore, we used the last SC (SC9) as an estimate of the least-scaling direction, and compared *Tp* to average MFC activity projected onto SC9. As predicted by the model, the average activity of the non-scaling components of MFC were indeed predictive of *Tp* (Supplementary Fig. 8). This is a compelling result as it bears out a key prediction about an unsuspected relationship between cortical activity and behavior made by a model that was constrained only to perform the task.

A potential neural mechanisms for speed control

To further investigate the role of input in speed control, we analyzed the eigenvalues of the system near $F_{terminal}$. In the vicinity of this fixed point, stronger inputs caused the eigenvalues to decrease systematically (Fig. 5f, left). In a linear dynamical system, such contraction in the eigenvalue spectrum corresponds to a systematic increase in the network's effective time constants, τ_{eff} (Fig. 5f, right). From this, we concluded that the action exerted by the input is equivalent to adjusting the system's effective time constant in a flexible input-dependent manner.

To gain insight into the mechanism that provides such powerful and modular control of time constants, we focused on a simplified model composed of only two mutually inhibitory neurons with a common input (Fig. 6a and Supplementary Note). Previous work has demonstrated that adjustments of the common input in this model could alter its recurrent dynamics to either relax to a single fixed point with a specific time constant or act as an integrator with exceedingly long time constants[38]. We reasoned that exploring the model's behavior while between these two regimes might lead us to a mechanistic understanding of how the effective time constant of a network can be flexibly adjusted.

In the presence of balanced input (Cue), the two-neuron model is associated with an energy landscape that engenders a pair of stable fixed points similar to the recurrent model (Fig. 6b). We analyzed the phase plane of the model (Fig. 6c) and verified that the input level can be used to create a continuum of τ_{eff}. This is analogous to the recurrent network model where activity along the input subspace served to control the speed. However, the two-neuron model helped us understand the underlying mechanisms: stronger input drives neurons toward their saturating nonlinearity where the slopes of activation functions are shallower (Fig. 6d). Shallower slopes correspond to smaller derivatives and larger values of τ_{eff}. In other words, the presence of single-neuron nonlinearities provides a reservoir of slopes that an input can exploit to control the network's energy gradients (Fig. 6b).

Having established a low-level mechanism in the two-neuron model, we asked whether the same mechanism was operative in the recurrent network model. For the recurrent model, we analyzed the operating point of units as a function of the input drive near $F_{terminal}$. Remarkably, for stronger inputs, units were systematically driven further toward their saturating nonlinearity (Fig. 5g, h), which is consistent with the mechanism of speed control in the simple network model. These results underscore a simple and powerful mechanism at the level of single neurons for controlling the speed of dynamics independent of the neural trajectory.

Discussion

We found that flexible motor timing was governed by controlling the speed of slow dynamics across populations of MFC and caudate neurons. Speed control also emerged as a natural solution in recurrent network models trained to produce different time intervals. This was achieved by an input that drove the system to the appropriate region of the state space where recurrent interactions unfolded at desired speeds. In both systems, fluctuations of speed predicted variability within each temporal context and systematic adjustments of

speed provided the means for flexible control of timing. These results suggest that the brain uses a speed control mechanism to deliberately control movement initiation time.

The division of labor conferred by the input and recurrent interactions has broad implications for tempo-flexible control of behavior allowing the same motor and cognitive functions to unfold along the same neural trajectory at different timescales. For example, in decision making tasks, adjustment of a speed command could explain how the brain might flexibly implement different speed-accuracy tradeoffs[39]. Indeed, if the speed command is controlled by a sensory input, our recurrent network would behave similar to more detailed network models consisting of excitatory and inhibitory units that approximate temporal integration of sensory information[40]. However, biophysical models of decision making have not yet been extended to generate the diversity of scaling response profiles that we observed *in-vivo* and in our recurrent model.

The engineered two-neuron model highlighted the crucial role of single-neuron nonlinearities; adjustments of speed were governed by the interaction of input with these nonlinearities. This finding suggests that circuits and subcircuits could exploit different inputs and different biophysical properties to adjust speed independently and operate at different timescales. It also predicts that neuromodulatory effects and pharmacological treatments that interfere with the nonlinear response curve of individual neurons could alter the speed of cortical dynamics, as observations from numerous studies of interval timing might suggest[41].

The source of the external input that adjusts the speed remains a pertinent and unresolved question. One possibility is that MFC received this input directly from neurons in other cortical areas, which is consistent with recent observation in the parietal cortex[42]. Another possibility is that the input has a thalamocortical origin. Thalamic neurons, in turn, may inherit this signal from other cortical and/or subcortical regions. Neuromodulatory signals could also alter cortical dynamics. A number of physiology and pharmacology studies have implicated dopamine in regulating timing behavior[43,44]. Cortical dynamics are also known to depend on cellular properties such as those mediated by NMDA receptors, which are thought to facilitate the generation of stable slow cortical dynamics[45].

Another question for future work concerns the exact mechanisms that give rise to the diversity of response profiles in MFC. According to our model, this diversity emerges from recurrent interactions in direct response to an input drive. Alternatively, these activity patterns could be the result of cortical nonlinearities acting upon simpler ramping inputs, which constituted a minority of response profiles in cortico-basal ganglia circuits we recorded from. Indeed, considering the bidirectional connections between thalamus and cortex, we cannot rule out the possibility that ramping activity in thalamus and/or other cortical areas might contribute to the scaling of more complex response profiles in MFC. Nevertheless, the model seems to provide the most parsimonious account of the data for both cortex and thalamus. The exact details of the signalling pathways, recurrent microcircuitry and biophysical properties notwithstanding, the mechanisms that we have identified have the potential to explain how the brain flexibly controls the speed of cortical dynamics.

Online Methods

Methods

Two adult rhesus monkeys (Macaca mulatta, a 6.5 kg female and 9.0 kg male, both 5 years old) were trained on a two-interval two-effector motor timing task. All surgical, behavioral and experimental procedures conformed to the guidelines of National Institutes of Health and were approved by the Committee of Animal Care at Massachusetts Institute of Technology.

Behavior

The MWorks software package (http://mworks-project.org) running on a Mac Pro was used to deliver stimuli and to control behavioral contingencies. Visual stimuli were presented on a 23 inch monitor at a refresh rate of 60 Hz. Eye positions were tracked with an infrared camera (Eyelink 1000; SR Research Ltd, Ontario, Canada) and sampled at 1 kHz. A custom-made manual button, equipped with a trigger and a force sensor, was used to register button presses.

Motor timing task—Each trial began with the appearance of two fixation cues (FCs), a circle at the center of the screen and a square 0.5 deg below the circle. The animal had to shift its gaze to the circle and the square informed the animal to hold its hand gently on the button. On each trial, one FC was colored and the other was white. The colored FC indicated the desired response effector (colored circle for saccade and colored square for button press). The color indicated the desired interval (red for 800 ms and blue for 1500 ms). We denote these four trial conditions by EL, ES, HL and HS where E and H refer to *Eye* and *Hand*, and S and L to *Short* (800 ms) and *Long* (1500 ms) intervals. After a delay period (500 – 1500 ms, uniform hazard), the saccade target was briefly presented 8 degrees to the left or right of the FC. For button press trials (colored square), the saccadic target was not relevant but was presented so that stimuli were consistent across trials. After another delay (500 – 1500 ms, uniform hazard), a 48 ms annulus (Set cue) flashed around the FCs cued the animal to start timing. Trials were aborted if the animal made premature eye or hand movements (before Set or long before the desired time). To receive reward, animals had to initiate a movement with the desired effector (cued by the colored FC) within a small window (“acceptance window”) around the desired interval (cued by the color of FC). The saccade responses had to land inside a circular window of radius 2.5 deg centered on the location of the extinguished target and had to be made directly (less than 33 ms after exiting the FC window). Button-press responses had to be made with the hand contralateral to the recorded hemifield[46]. The production interval was measured from the endpoint of Set to the moment the saccade was initiated or the button was triggered. The width of the acceptance window was adjusted dynamically on a trial-by-trial basis and independently for the *Short* and *Long* conditions using a one-up one-down staircase procedure. As such, animals were rewarded for nearly half of trials (on average, 57% in monkey A and 51% in monkey D) for both temporal contexts. For trials that were rewarded, in addition to reward delivery, the color of the stimulus changed to green and an auditory clicking sound was simultaneously presented. Within the acceptance window, the magnitude of the reward scaled with accuracy.

Electrophysiology

Animals were comfortably seated in a dark and quiet room. Each session began with an approximately 10-minute warmup period to allow animals to recalibrate their timing and exhibit stable behavior during electrophysiology recordings. Recordings were made through a craniotomy within a recording chamber while the animal's head was immobilized. Structural MRI scans were used to aid in targeting regions of interest. Single- and multi-unit responses were recorded using a 24-channel laminar probe with 100 μm or 200 μm interelectrode spacing (V-probe, Plexon Inc.). Eye position was sampled at 1 kHz, and all behavioral and electrophysiological data were timestamped at 30 kHz and streamed to a data acquisition system (OpenEphys).

The dataset collected for this study included 1967 single- or multi-units recorded from the MFC, caudate and thalamus of two monkeys (Table 2), in which 69% (1351/1967) were tentatively single units. Neurons with firing rates less than 2 spikes per second during the timing epoch were excluded from subsequent analyses.

Reversible inactivation

Injections were made with a microinjection pump (UMP3, World Precision Instruments) and a Hamilton syringe, which was connected to a custom 30G stainless steel injection cannula via a fused silica injection line (365μm OD, 100μm ID, Polymicro Technologies). In each injection session, we first established the animal's baseline behavioral performance. Afterwards, we pressure-injected muscimol hydrobromide (5 μg/μL in saline) in the region of interest at a rate of 0.2 μL/min. In the MFC and caudate, a total of 2 μL was injected per session. In pilot inactivation experiments in the thalamus, we noticed that animals stopped performing the task after 2 μL muscimol injection. To ensure animals would perform the task, the total volume of muscimol in the thalamus was reduced to 1.5 μL. The behavioral task was resumed 10 min after the the injection was completed. As a control, in separate sessions, sterile saline was injected following the same procedure. The experimental data consisted of unequal test sessions for muscimol and saline, and unequal number of trials in the before and after muscimol injection. For statistical comparison, these inequalities may introduce sampling biases. In order to avoid such biases, we created 50-trial mini-sessions from before and after the injections in which the trials within a mini-session were randomly sampled. The sampling was made without repeats to ensure trials were not counted twice. We quantified the effects of inactivation by comparing mean squared error, Bias and Variance $MSE = \Sigma(T_p - T_s)^2 = Bias^2 + Var$ before and after the injection for every mini-session. The same procedure was used to assess results of the saline injection experiments.

Antidromic Stimulation

We used antidromic stimulation to localize thalamocortical MFC-projecting neurons. Antidromic spikes were recorded on a 24-channel electrode (V-probe, Plexon Inc.) in response to a single biphasic pulse of duration 0.2 ms (current < 500 uA) delivered to MFC via low impedance tungsten microelectrodes (100 – 500 KΩ, Microprobes). The guide tube for the tungsten electrode was used as the return path for the stimulation current. Antidromic activation evoked spikes reliably at a latency ranging from 1.8 to 3 ms, with less than 0.2 ms

jitter. The region of interest targeted in the thalamus was within 1 mm of antidromically identified neurons.

Mathematical notation

Throughout the manuscript, we have used lowercase for scalars (x), bold and lowercase for vectors ($\boldsymbol{x}$), bold and uppercase for matrices (X). Brackets were used for indexing values within vectors and matrices ($\boldsymbol{x}[i]$ and $\boldsymbol{X}[i, j]$). Subscripts were used for indexing a set of scalars (x_j), vectors ($\boldsymbol{x}_j$), or matrices ($\boldsymbol{X}_j$). Superscripts were also used to show projections onto a subspace. For example, $\boldsymbol{x}_{PC(1:k)}$ refers to a vector projected onto a the first k principal components. Curly brackets were used to indicate a subset of conditions. For example, $\boldsymbol{x}\{a = a_0; b = b_1\}$ refers to a vector computed for a subset of trials in which both $a = a_0$ and $b = b_1$ conditions were satisfied. The symbol ∪ is used to indicate data combined across a number of variables. For example $\cup_i^N \{\boldsymbol{x}_i\}$ denotes data collected across a union of vectors $\boldsymbol{x}_i$. The symbol $< \boldsymbol{x} >_i$ was used to show averaging of a vector $\boldsymbol{x}$ across i. Point functions were shown as lowercase ($f(.)$) regardless of whether they were applied to scalars or vectors.

Data Analysis

All offline data processing and analyses were performed in MATLAB (2016b, MathWorks). Spiking data were bandpass filtered between 300 Hz to 7 kHz and spike waveforms were detected at a threshold that was typically set to 3 times the RMS noise. Single- and multi-units were sorted offline using a custom software, MKsort (https://github.com/ripple-neuro/mksort). The majority of the neurons were recorded in separate behavior sessions.

Estimating firing rates accurately is challenging when rates change dynamically and trials have different durations[47,48], which was the case in our data. Since our focus was on firing rates leading up to the movement, we aligned trials with respect to movement time (Fig. 2c). Additionally, for each condition, we discarded trials with Tps that were more than 3 standard deviations away from the mean (1.46% of trials). Firing rates were estimated by: 1) averaging spike counts per time bin, 2) using a 40 ms Gaussian kernel to compute smooth spiking density functions, and 3) z-scoring to minimize sampling bias due to baseline and amplitude differences across neurons.

To examine the relationship between firing rates and Tps, we binned trials according to Tp and compared average firing rates for each bin. For the 800 ms interval, we used 7 bins centered on 740 to 860 ms every 20 ms, and for the 1500 ms, we used 9 bins centered on 1300 to 1620 ms every 40 ms. We denoted the average firing rate of a neuron as a function of time by $r(t)$, average firing rate for a specific condition c (EL, ES, HL or HS) by $r(t;c)$, and average firing rate for a specific condition and a specific Tp bin by $r(t;c, Tp)$. For population analyses, response vectors of individual neurons were organized into rows of a matrix denoted by $\boldsymbol{r}(t;c, Tp)$.

To test if activity profiles could be described by a linear function (e.g. ramping activity), we compared 0 to 8th order polynomial fits to $r(t)$ using cross-validation with randomized train and test sets. All neurons that were best explained by a polynomial of order 0 or 1 were considered linear so long as the fit explained at least 50% of variance. We also applied the

same procedure allowing up to 200 ms offset from the beginning or end of the timing interval to ensure our results were robust.

Compare the motor timing models at the level of single/multi-units

To avoid overfitting and facilitate comparison of models with different levels of complexity, all model fitting was performed on the training set and the goodness of fit (R^2) was quantified on the test set. In the clock-accumulator model with a flexible threshold, a linear ramp with fixed slope and different thresholds for different production intervals was fit to the response profile. In the clock-accumulator model with a flexible clock, the threshold was fixed and ramping rate was adjusted according to the interval. In the clock-accumulator model with both flexible clock and flexible threshold, a linear ramp was adjusted according to the interval and its offset was independently adjusted for each interval. In the oscillation based models, sinusoidal functions or a sum of up to 4 different sinusoids were fit to activity profiles, in which the frequency, amplitude and phase for each sinusoid were free parameters. In the population clock, a single polynomial of up to 8th order was fit to the response profiles for both *Short* and *Long* contexts. For the temporal scaling model, the response profiles for the *Short* condition was used to find the best-fitting polynomial, and the temporally scaled version of the fitted functions were used to test the goodness of fit for *Long* trials.

Scaling Subspace

We used a principal component analysis (PCA) as a first step to compute a low-dimensional and unbiased estimate of data. We found that the first 9 principal components (PCs) captured nearly 80% of the variance in the data (Fig. 3b, bottom). We therefore computed the scaling components (SCs) from data captured by the first 9 PCs, which was computed as follows:

$\boldsymbol{r}_{PC}(t;c) = \hat{\boldsymbol{V}} * \boldsymbol{r}(t;c)$ and $\hat{V} = [\upsilon_1^T; \upsilon_1^T; \ldots; \upsilon_{N_{\mathrm{PC}}}^T]$ is the projection matrix, in which υ_i is the i^{th} PC direction. Therefore, the denoised activity across all conditions and time points $\boldsymbol{r}_{PC}(t;c)$ is of size $N_{pc} \times (T \times C)$. We computed the corresponding scaled responses using our scaling procedure and denoted the result by $r_{\mathrm{PC}}^S(t;c)$. To find the scaling subspace, we solved an optimization problem that minimized the difference between average firing rates associated with different *Tps* (e.g., Tp_i and Tp_j). We denote the corresponding projection by U_{SC} and refer to its columns as scaling components (SCs). The resulting projection $\boldsymbol{r}_{SC}$ can be computed as follows:

$$U_{\mathrm{SC}} = \arg\min_{U} \{\mathrm{var}(U^*[\boldsymbol{r}_{\mathrm{PC}}^S(t;\mathrm{Tp}_i) - \boldsymbol{r}_{\mathrm{PC}}^S(t;\mathrm{Tp}_j)])\}$$

$$\boldsymbol{r}_{\mathrm{SC}}(t;c) = U_{\mathrm{SC}} * \boldsymbol{r}_{\mathrm{PC}}(t;c)$$

We hypothesized that the speed of activity in the scaling subspace predicts *Tp*. We computed the instantaneous speed in the scaling subspace from projections of responses on to the first three SCs as follows:

$$S(T_p)=\frac{1}{T}\sum\nolimits_{t=1:T}\|d\boldsymbol{r}_{\mathrm{SC}1:3}(t,T_p)/\mathrm{dt}\|$$

$$\boldsymbol{r}_{\mathrm{SC}1:3}(t;c)=U_{\mathrm{SC}1:3}*\boldsymbol{r}_{\mathrm{PC}}(t;c)$$

For each interval bin, we obtained an unbiased estimate of the relationship between speed and *Tp* by resampling trials with replacement within each interval bin. The relationship between the average speed *S*(*Tp*) and production intervals was fitted in the log space by a linear function:

$$\log(S(\mathrm{Tp}))=A-B.\log(\mathrm{Tp})$$

Scaling index for population data

We quantified temporal scaling in single units, principal components (PCs) and scaling components (SCs) using a scaling index (*SI*) that represented a general measure of the degree of similarity between multiple response profiles associated with different intervals. *SI* was computed as follows: (1) trials were sorted based on production interval (*Tp*); (2) sorted trials were grouped into bins of similar *Tp*s (as described previously in Methods); (3) the first 9 PCs and the corresponding SCs for each bin were computed; (4) for each PC and SC, the index was computed as the coefficient of determination (R^2) after the PCs and SCs were temporally scaled. This metric, which varies between 0 and 1, quantifies the degree to which each PC/SC undergoes temporal scaling for different *Tp*s.

$$z_{\mathrm{scaled}}=\boldsymbol{r}^{s}_{\mathrm{PC/SC}}(t;\cup_i^N\{\boldsymbol{Tp}_i\})$$

$$\mathrm{SI}=\frac{\sum_{t=1}^{N_t}[z_{\mathrm{scaled}}-<z_{\mathrm{scaled}}>_{\mathrm{Tp}}]^2}{\sum_{t=1}^{N_t}\sum_{\mathrm{Tp}=\mathrm{Tp}^1}^{\mathrm{Tp}^n}[z_{\mathrm{scaled}}-<z_{\mathrm{scaled}}>_{\mathrm{Tp},t}]^2}$$

We evaluated the degree of scaling among populations in each region of interest by computing the scaling index for each PC and SC in those populations. Additionally, we computed the variance explained by each SC. Finally, to gain an unbiased estimate of the relationship between variance explained and scaling index, we computed these two metrics along randomly selected dimensions within the state space. This analysis revealed the full distributions of variance explained and scaling index and their relationship within the whole state space.

Recurrent Network Architecture

We constructed a firing rate recurrent neural network (RNN) model with *N* nonlinear units (*N* = 200). The network dynamics was governed by the following differential equation:

$$\boldsymbol{F}(\boldsymbol{x})=\boldsymbol{\tau}\dot{\boldsymbol{x}}(t)=-\boldsymbol{x}+\boldsymbol{J}\boldsymbol{r}(t)+\boldsymbol{B}\boldsymbol{u}+\boldsymbol{c}_{\boldsymbol{x}}+\boldsymbol{\rho}(t)$$

$$\boldsymbol{r}(t)=\tanh(\boldsymbol{x}(t))$$

Variable $\boldsymbol{x}(t)$ is an N-dimensional vector representing the activity of all the units. Variable $\boldsymbol{r}(t)$ represents the firing rates of those units by transforming x through a *tanh* saturating nonlinearity. The time constant of each neuron was set to $\boldsymbol{\tau} = 10ms$ This value is different from $\boldsymbol{\tau}_{eff}$, which emerges at the network-level. Variable $\boldsymbol{c}_{\boldsymbol{x}}$ is a vector representing a stationary offset the units receive, and $\boldsymbol{\rho}(t)$ is a vector representing white noise $N(0,0.1)$ sampled at each timestep $\Delta t = 1ms$. The recurrent connections in the network are specified by matrix $\boldsymbol{J}$, whose values, following previous work on balanced networks, are drawn from a normal distribution with zero mean and variance $1/N$. The network receives a two-dimensional input u consisting of a context cue $u_c(t)$ and a transient Set pulse $u_s(t)$. The network received these inputs through synaptic weights $\boldsymbol{B}=[b_c, b_s]$, which were initialized to random values drawn from a uniform distribution with range −1 to 1.

The context input, u_c, represents the interval-dependent context cue input (color). The value of u_c was was set to zero for 100 ms and then jumped to a graded value proportional to the length of one of 16 desired intervals distributed within a range 500 – 1700 ms. The offset of u_c was sampled proportionally from the range 0.1 to 0.6 and was perturbed with Gaussian noise $N(0,.025)$ at each Δt. Increasing input noise did not qualitatively alter the network training solutions. The transient 'Set' pulse $u_s(t)$ was active for 10 ms with magnitude 0.1 and zero elsewhere. On each training and test trial, the interval between the onset of u_c and $u_s(t)$ was drawn from a uniform distribution with range (100 – 200 ms).

The network produced a one-dimensional output $z(t)$, read-out by the summation of linear units with weights $\boldsymbol{w}_o$ and a bias term c_z. The output weights were initialized to zero at the start of training.

$$z(t)=\boldsymbol{w}_o^T\boldsymbol{r}(t)+c_z$$

Statistics

The weber fractions across behavioral sessions (Fig. 1c), MSEs before and after inactivation (across mini-sessions, Table 1 and Figs. 2b, an 4b,f), scaling indices obtained from a bootstrap procedure for various brain areas and surrogate data (Fig. 3e) were assumed to be normally distributed, however this was not formally tested a priori. Depending on assumptions associated with various sessions, one-tailed paired or unpaired sample t-tests were used. Neurons with extremely low firing rates (less than 2 spk/sec) during the timing epoch were excluded from further analysis. The number of neurons recorded in all three areas in both monkeys and those excluded have been reported in Table 2. For the single neuron responses with respect to the seven types of timing models, we used a one-way ANOVA test to establish that the explanatory power quantified by R^2 of various models were

significantly different. Then, we used post-hoc paired-sample t-tests to compare temporal scaling model with each alternative model (Fig. 2e). The scaling indices of neurons in different brain areas (Supplementary Fig. 1a) were not normally distributed. For this reason, we used a nonparametric unpaired Mann-Whitney-Wilcoxon test to compare independent samples from the two brain areas under examination (thalamus and MFC, thalamus and caudate). Please see additional information in the Life Sciences Reporting Summary.

Data Availability statement and Accession Code Availability Statements

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank M.S. Fee, J.J. DiCarlo, and R. Desimone for comments on the manuscript, and David Sussillo for advice on modeling. D. Narain was supported by the Rubicon Grant (2015/446-14-008) from the Netherlands Scientific Organization (NWO). M. Jazayeri is supported by NIH (NINDS-NS078127), the Sloan Foundation, the Klingenstein Foundation, the Simons Foundation, the Center for Sensorimotor Neural Engineering, and the McGovern Institute.

References

1. Stuphorn V, Schall JD. Executive control of countermanding saccades by the supplementary eye field. Nat. Neurosci. 2006; 9:925–931. [PubMed: 16732274]
2. Kunimatsu J, Tanaka M. Alteration of the timing of self-initiated but not reactive saccades by electrical stimulation in the supplementary eye field. Eur. J. Neurosci. 2012; 36:3258–3268. [PubMed: 22845785]
3. Fried I, et al. Functional organization of human supplementary motor cortex studied by electrical stimulation. J. Neurosci. 1991; 11:3656–3666. [PubMed: 1941101]
4. Lewis PA, Wing AM, Pope PA, Praamstra P, Miall RC. Brain activity correlates differentially with increasing temporal complexity of rhythms during initialisation, synchronisation, and continuation phases of paced finger tapping. Neuropsychologia. 2004; 42:1301–1312. [PubMed: 15193939]
5. Shima K, Tanji J. Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. J. Neurophysiol. 2000; 84:2148–2160. [PubMed: 11024102]
6. Isoda M, Hikosaka O. Switching from automatic to controlled action by monkey medial frontal cortex. Nat. Neurosci. 2007; 10:240–248. [PubMed: 17237780]
7. Lu X, Matsuzawa M, Hikosaka O. A neural correlate of oculomotor sequences in supplementary eye field. Neuron. 2002; 34:317–325. [PubMed: 11970872]
8. Mello GBM, Soares S, Paton JJ. A scalable population code for time in the striatum. Curr. Biol. 2015; 25:1113–1122. [PubMed: 25913405]
9. Gouvêa TS, et al. Striatal dynamics explain duration judgments. Elife. 2015; 4
10. Jin DZ, Fujii N, Graybiel AM. Neural representation of time in cortico-basal ganglia circuits. Proceedings of the National Academy of Sciences. 2009; 106:19156–19161.
11. Matell MS, Meck WH, Nicolelis MAL. Interval timing and the encoding of signal duration by ensembles of cortical and striatal neurons. Behav. Neurosci. 2003; 117:760–773. [PubMed: 12931961]
12. Merchant H, Harrington DL, Meck WH. Neural basis of the perception and estimation of time. Annu. Rev. Neurosci. 2013; 36:313–336. [PubMed: 23725000]

13. Mauk MD, Buonomano DV. The neural basis of temporal processing. Annu. Rev. Neurosci. 2004; 27:307–340. [PubMed: 15217335]
14. Bartolo R, Prado L, Merchant H. Information processing in the primate basal ganglia during sensory-guided and internally driven rhythmic tapping. J. Neurosci. 2014; 34:3910–3923. [PubMed: 24623769]
15. Schultz W, Romo R. Neuronal activity in the monkey striatum during the initiation of movements. Exp. Brain Res. 1988; 71:431–436. [PubMed: 3169174]
16. Tanaka M. Inactivation of the central thalamus delays self-timed saccades. Nat. Neurosci. 2006; 9:20–22. [PubMed: 16341209]
17. Treisman M. Temporal discrimination and the indifference interval. Implications for a model of the 'internal clock'. Psychol. Monogr. 1963; 77:1–31.
18. Killeen PR, Fetterman JG. A behavioral theory of timing. Psychol. Rev. 1988; 95:274–295. [PubMed: 3375401]
19. Matell MS, Meck WH. Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. Brain Res. Cogn. Brain Res. 2004; 21:139–170. [PubMed: 15464348]
20. Karmarkar UR, Buonomano DV. Timing in the absence of clocks: encoding time in neural network states. Neuron. 2007; 53:427–438. [PubMed: 17270738]
21. Buonomano DV, Laje R. Population clocks: motor timing with neural dynamics. Trends Cogn. Sci. 2010; 14:520–527. [PubMed: 20889368]
22. Brody CD, Hernández A, Zainos A, Romo R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb. Cortex. 2003; 13:1196–1207. [PubMed: 14576211]
23. Komura Y, et al. Retrospective and prospective coding for predicted reward in the sensory thalamus. Nature. 2001; 412:546–549. [PubMed: 11484055]
24. Merchant H, Zarco W, Pérez O, Prado L, Bartolo R. Measuring time with different neural chronometers during a synchronization-continuation task. Proc. Natl. Acad. Sci. U. S. A. 2011; 108:19784–19789. [PubMed: 22106292]
25. Emmons EB, et al. Rodent Medial Frontal Control of Temporal Processing in the Dorsomedial Striatum. J. Neurosci. 2017; 37:8718–8733. [PubMed: 28821670]
26. Gibbon J. Scalar expectancy theory and Weber's law in animal timing. Psychol. Rev. 1977; 84:279.
27. Rakitin B, et al. Scalar expectancy theory and peak-interval timing in humans. J. Exp. Psychol. Anim. Behav. Process. 1998; 24:15–33. [PubMed: 9438963]
28. Douglas Creelman C. Human Discrimination of Auditory Duration. J. Acoust. Soc. Am. 1962; 34:582–593.
29. Gibbon J, Church RM, Meck WH. Scalar timing in memory. Ann. N. Y. Acad. Sci. 1984; 423:52–77. [PubMed: 6588812]
30. Grondin S. From physical time to the first and second moments of psychological time. Psychol. Bull. 2001; 127:22–44. [PubMed: 11271754]
31. Hikosaka O, Sakamoto M, Usui S. Functional properties of monkey caudate neurons. I. Activities related to saccadic eye movements. J. Neurophysiol. 1989; 61:780–798. [PubMed: 2723720]
32. Parthasarathy HB, Schall JD, Graybiel AM. Distributed but convergent ordering of corticostriatal projections: analysis of the frontal eye field and the supplementary eye field in the macaque monkey. J. Neurosci. 1992; 12:4468–4488. [PubMed: 1279139]
33. Sussillo D, Churchland MM, Kaufman MT, Shenoy KV. A neural network that finds a naturalistic solution for the production of muscle activity. Nat. Neurosci. 2015; 18:1025–1033. [PubMed: 26075643]
34. Li N, Daie K, Svoboda K, Druckmann S. Robust neuronal dynamics in premotor cortex during motor planning. Nature. 2016; 532:459–464. [PubMed: 27074502]
35. Mante V, Sussillo D, Shenoy KV, Newsome WT. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature. 2013; 503:78–84. [PubMed: 24201281]
36. Hanes DP, Schall JD. Neural control of voluntary movement initiation. Science. 1996; 274:427–430. [PubMed: 8832893]

37. Sussillo D, Barak O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Comput. 2013; 25:626–649. [PubMed: 23272922]
38. Machens CK, Romo R, Brody CD. Flexible control of mutual inhibition: a neural model of two-interval discrimination. Science. 2005; 307:1121–1124. [PubMed: 15718474]
39. Hanks T, Kiani R, Shadlen MN. A neural mechanism of speed-accuracy tradeoff in macaque area LIP. Elife. 2014; 3
40. Wang X-J. Probabilistic decision making by slow reverberation in cortical circuits. Neuron. 2002; 36:955–968. [PubMed: 12467598]
41. Meck W. Neuropharmacology of timing and time perception. Brain Res. Cogn. Brain Res. 1996; 3:233.
42. Jazayeri M, Shadlen MN. A Neural Mechanism for Sensing and Reproducing a Time Interval. Curr. Biol. 2015; 25:2599–2609. [PubMed: 26455307]
43. Buhusi CV, Meck WH. Differential effects of methamphetamine and haloperidol on the control of an internal clock. Behav. Neurosci. 2002; 116:291–297. [PubMed: 11996314]
44. Soares S, Atallah BV, Paton JJ. Midbrain dopamine neurons control judgment of time. Science. 2016; 354:1273–1277. [PubMed: 27940870]
45. Chaudhuri R, Knoblauch K, Gariel M-A, Kennedy H, Wang X-J. A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. Neuron. 2015; 88:419–431. [PubMed: 26439530]
46. Fujii N, Mushiake H, Tanji J. Distribution of eye- and arm-movement-related neuronal activity in the SEF and in the SMA and Pre-SMA of monkeys. J. Neurophysiol. 2002; 87:2158–2166. [PubMed: 11929933]
47. K Namboodiri VM, Namboodiri VMK, Hussain Shuler MG. Report of interval timing or action? Proceedings of the National Academy of Sciences. 2014; 111:E2239–E2239.
48. Xu M, Zhang S-Y, Dan Y, Poo M-M. Representation of interval timing by temporally scalable firing patterns in rat prefrontal cortex. Proc. Natl. Acad. Sci. U. S. A. 2013; :1–6. DOI: 10.1073/pnas.1321314111

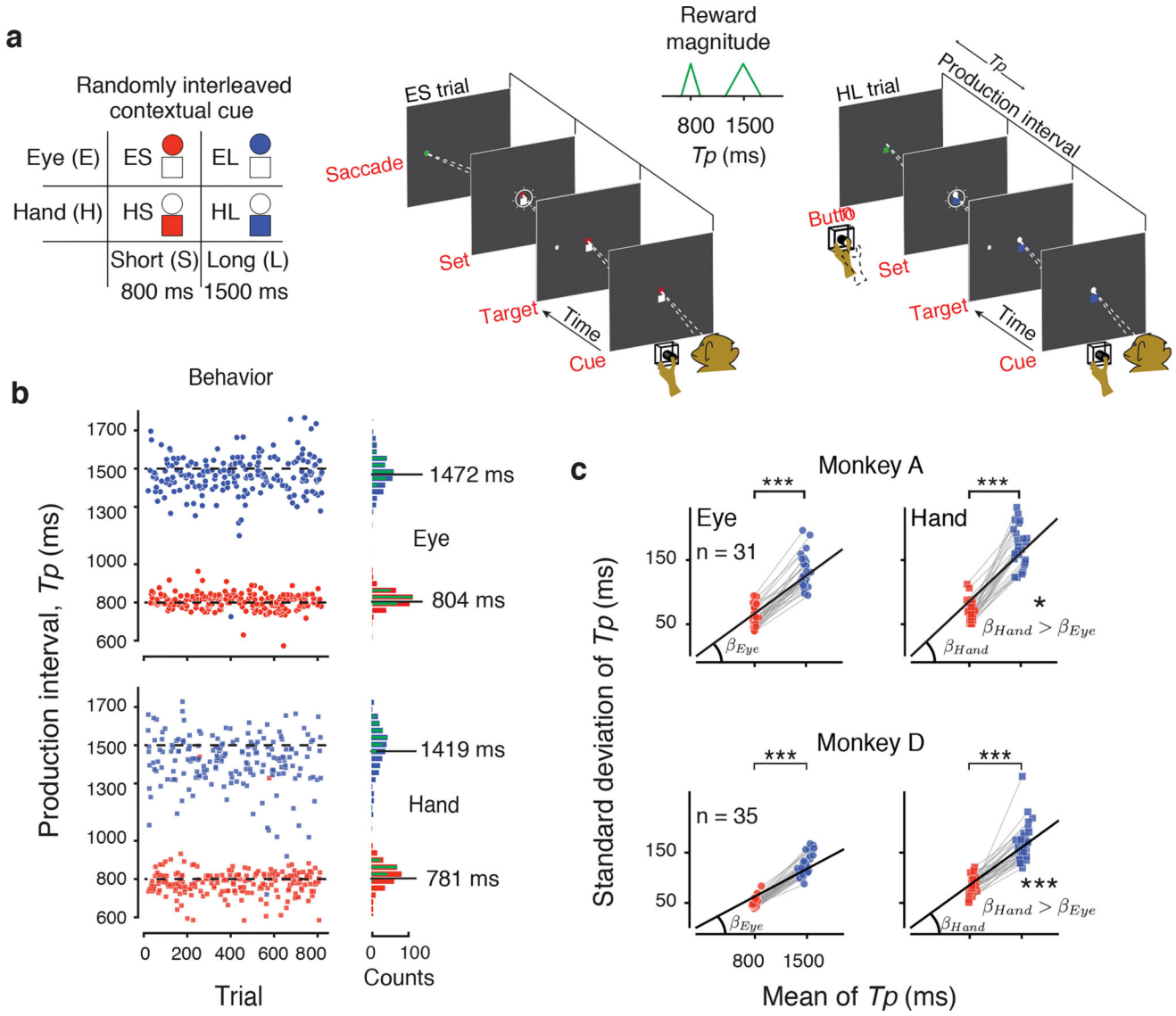

Fig. 1. Time production task and behavior

(a) Trial structure. Animals produced either an 800 ms (*Short*) or a 1500 ms (*Long*), either by making a saccade (*Eye*) or a button press (*Hand*). These four conditions were randomly interleaved and were cued throughout the trial by the color and shape of two central stimuli, a circular fixation for the eye and a square that cued the animal to place its hand on a button. The colored shape (circle or square) cued the effector, and the hue (red or blue) cued the desired interval (red for *Short* and blue for *Long*). After a random delay, a white circle was flashed to the left or right of the fixation point. This peripheral flash specified the saccadic target for the eye trials and played no role in the hand trials. After another random delay, a *Set* cue (a ring flashed around the fixation stimuli) initiated the motor timing epoch. The animal's production interval (*Tp*) was measured as the interval between *Set* and when either the saccade was made or the button was pressed. When *Tp* was generated with the desired effector and was within a specified reward window, the peripheral target (or the square fixation) turned green, auditory feedback was provided, and animal received juice. The

reward window was adjusted adaptively on a trial-by-trial basis and independently for the *Short* and *Long* conditions so that the animal received reward on approximately 50% of trials for both interval context on every session (on average, 57% in monkey A and 51% in monkey D). The reward magnitude increased linearly with accuracy as shown by the green triangular reward function. Two example trials, one for the *Eye*/*Short* (ES) condition (left) and one for the *Hand*/*Long* (HL) condition are shown. **(b)** A typical behavioral session showing *Tp* while the animal flexibly switched between the four trial conditions. For clarity the *Eye* (left) and *Hand* (right) trials are plotted separately although during the task they were randomly interleaved. The top 4 histograms show the distribution of *Tp* for each condition with rewarded trials in green. The vertical lines correspond to the mean values that are also reported numerically. **(c)** For both effectors (left: *Eye*, right: *Hand*) and both animals (top: animal A, bottom: animal D), the standard deviation of *Tp* scaled with mean *Tp* (red: *Short*, blue: *Long*). For animal A, the mean ± s.e.m of *Tp*s across the conditions were ES: 810±48.9 ms, EL: 1495±117 ms, HS: 822.3±53.7 ms, HL: 1486±136 ms. For animal D, they were ES: 808±56.1 ms, EL: 1481±137 ms, HS: 836.7±91.3 ms HL: 1521±177 ms. The variability was significantly higher for the *Long* compared to the *Short*. The average Weber fraction (ratio of standard deviation to mean) for the *Hand* (β_{Hand}) was significantly larger that *Eye* (β_{Eye}) (one-tailed paired sample t-test, for monkey A, n = 31, $t_{30} = 1.80$, $P = .041$, and for monkey D, n = 35, $t_{34} = 6.44$, $P < .001$).

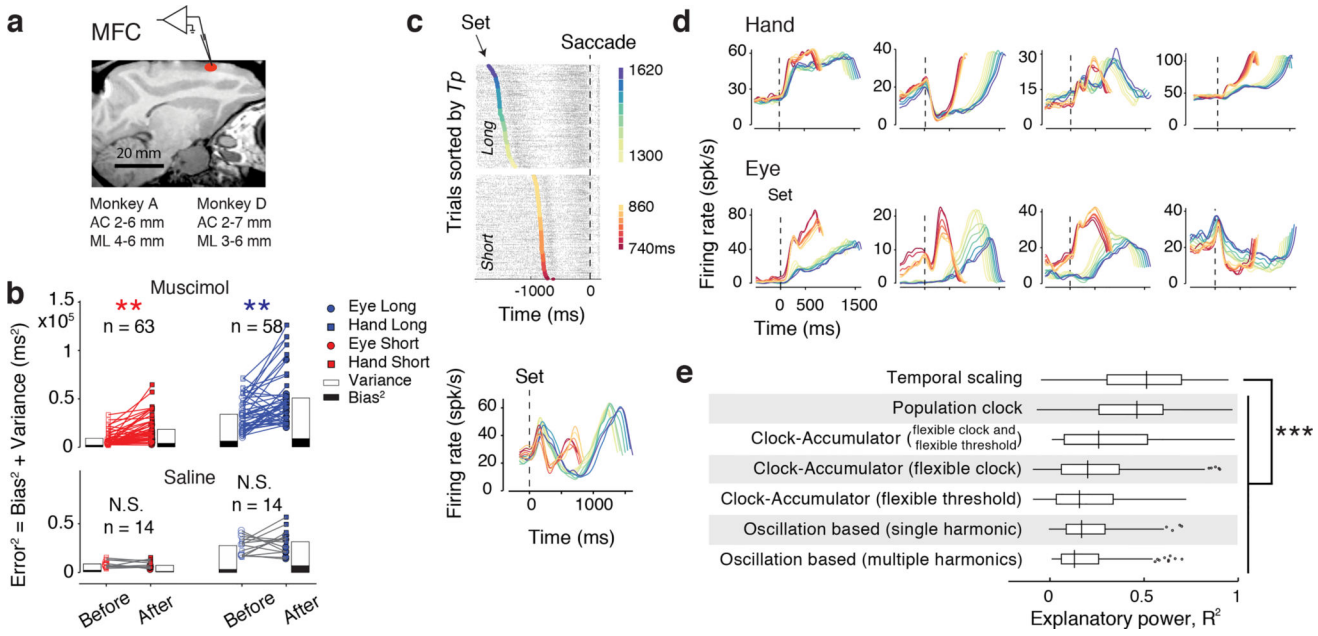

Fig. 2. Medial frontal cortex inactivation and electrophysiology

(a) Parasagittal view of the brain of one animal (monkey D) with a red ellipse showing the targeted region. Stereotactic coordinates used in each animal are shown with respect to anterior commissure (AC) and midline (ML). **(b)** Muscimol inactivation. Each line in each panel shows the change in mean squared error (MSE = $\Sigma(Tp - Ts)^2$ = Bias^2 + Var) computed from mini-session (randomly sampled subsets of trials without replacement; see Methods) before and after the injection of muscimol (above) and saline (below) for the two intervals (red: *Short*, blue: *Long*) and two effectors (circle: *Eye*, square: *Hand*). The white-over-black bar graphs partition MSE to Bias^2 (black) and Variance (white). Significance tests correspond to comparisons of MSE (see Table 1 for details) across mini-sessions (n: number of mini-sessions, **: $P < .001$, N.S.: not significant). **(c)** Average firing rates were computed after aligning spike times to movement initiation time. Top: Raster plot of spike times (black ticks) for an example neuron aligned to movement initiation time (dashed line) across trials (rows). Trials were sorted and grouped into bins according to the produced interval (*Tp*). Bottom: Average firing rates for each *Tp* bin plotted with respect to the time of Set (dashed line). The Set time in the top panel, and the activity profiles in the bottom panels were colored according to *Tp* bins (legend). **(d)** Activity profile of 8 example neurons for *Hand* (top) and *Eye* (bottom) conditions computed as described in (c). **(e)** Analysis of single neurons with respect to various model of timing (n = 416 neurons for both animals). Whisker plot showing the range of R^2 values captured by seven models fitted to the average firing rates of individual neurons (median: center line; box: 25th to 75th percentiles; whiskers: ±1.5× the interquartile range; dots: neurons whose R^2 values lie outside whiskers). The "Temporal scaling" model (top) had the highest explanatory power (R^2) across models (one-way ANOVA, $F_{6, 2859} = 125.2$, $P < .001$, and one-tailed paired sample *t*-test between 'Temporal scaling' and 'Population clock' model, n = 416, $t_{415} = 6.32$, ***$P < .001$). Models firs were cross-validated.

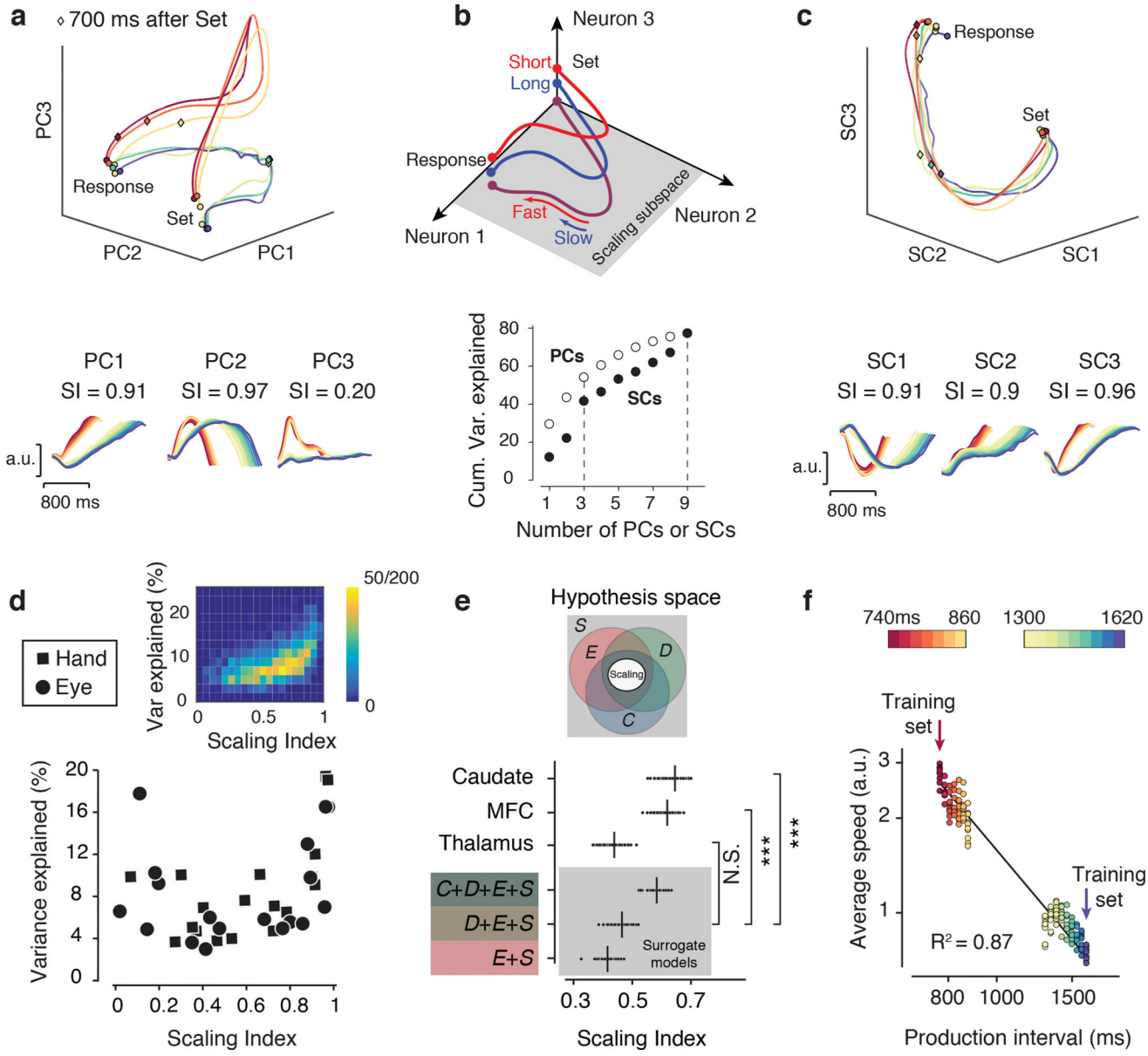

Fig. 3. Temporal scaling in the medial frontal cortex at the population level

(a) Top: Population activity for *Hand* trials for Monkey A projected onto the first 3 principal components (PCs) from the time of Set to the time of button press (Response). Activity profiles associated with different produced intervals are plotted in different colors (see color bar in Fig. 3f). Diamond shows activity 700 ms after Set. Bottom: The time course of the first three PCs with the corresponding scaling index (SI) values. **(b)** Top: Schematic drawing illustrating the scaling subspace. The response dynamics associated with *Short* (red) and *Long* (blue) produced interval (*Tp*) are depicted as distinct trajectories in the state space. Projections of neural responses onto a scaling subspace result in overlapping trajectories (purple) whose speed determines the produced interval, fast for *Short* (red) and slow for *Long* (blue). Bottom: Cumulative percentage variance explained by PCs and scaling components (SCs). **(c)** Top: Population activity sorted according to *Tp* bins and projected

onto the first 3 SCs. As expected, in this subspace, the trajectories overlap. Bottom: The first three SCs with the corresponding SI values. Because of cross-validation, SIs were not in decreasing order (see text). **(d)** Variance explained for individual SCs as a function of SI. SCs with the larger SI explain a large percentage of variance for both *Hand* (square) and *Eye* (circle) conditions. Inset: Variance explained as a function of SI derived along 200 random one-dimensional projections of MFC activity in the state space. Individual projections were binned and pseudocolored to indicate the frequency of occurrence. The data shows that high scaling indices are associated with high variance explained. **(e)** Comparison of SI in the MFC, caudate and thalamus with surrogate data generated from three Gaussian process models that were constrained to match the observed response profiles with increasing levels of sophistication (Supplementary Note and Supplementary Fig. 3). The inset shows the hypothesis space in relation to various constraints and their combinations with distinct colors and their overlaps. Perfect scaling (middle ellipse) is a subset of the possibilities that satisfy all four constraints. Each model consisted of the same number of neurons as that in the MFC data, and the number of bootstrapped samples for each model was n = 200. The plot shows the the average SI across all SCs computed from bootstraps (small circles) along with the corresponding mean (vertical line) for each of three brain areas and each of the surrogate models. The average SI for each surrogate model was significantly lower than the values associated with the MFC and caudate, but not for the thalamus (see main text for statistics). **(f)** The speed of neural trajectory within the scaling subspace spanned by the first 3 SCs predicted average *Tp*s across bins. The relationship between speed and *Tp* was fit to a linear log-log function. The scaling subspace was computed from training data (arrows, 2 *Tp* bins) and used to evaluate speed on the remaining test data (14 *Tp* bins). R^2 was computed by repeating the procedure using bootstrapping (n = 10). Both axes are in log scale.

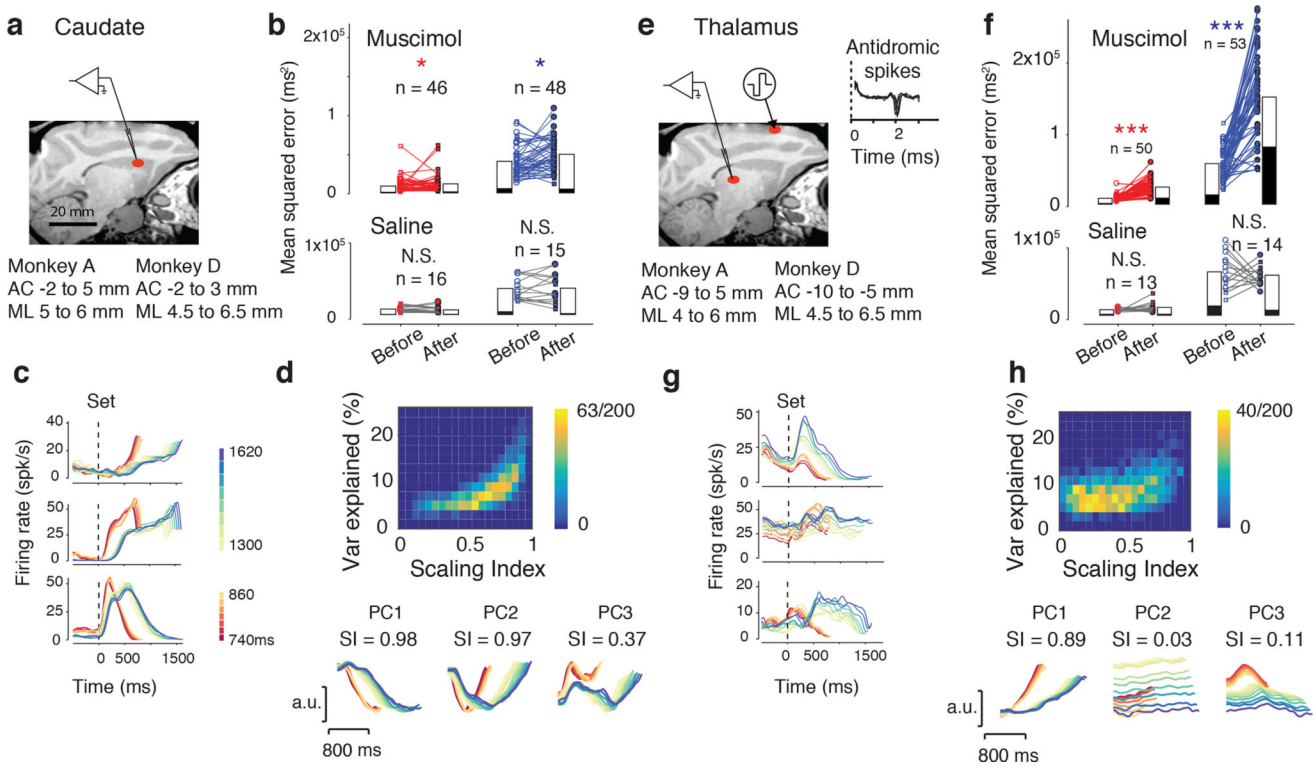

Fig. 4. Inactivation, electrophysiology and temporal scaling in the caudate and thalamus

(a) Same as Fig. 2a with a red ellipse and stereotactic coordinates showing targeted regions in the caudate. **(b)** Muscimol inactivation in the caudate. Results are presented in the same format as in Fig. 2b. **(c)** Activity profile of three example caudate neurons (same format as in Fig. 2d). **(d)** Top: The relationship between variance explained and scaling index (SI) in the caudate (same format as the inset of Fig. 3d). Bottom: The first three PCs with the corresponding SI values. **(e)** Same as panel a showing the region of interest in the thalamus. We recorded from neurons in the region where MFC-projecting neurons were identified antidromically. Inset: example of reliable and low-latency spikes detected after antidromic stimulation. **(f–h)** Inactivation, electrophysiology and temporal scaling in the thalamus (same format as panels b–d). Responses in the thalamus are qualitatively different from the caudate (panel d) and MFC (Fig. 3d) in that most projections in the state space do not exhibit temporal scaling.

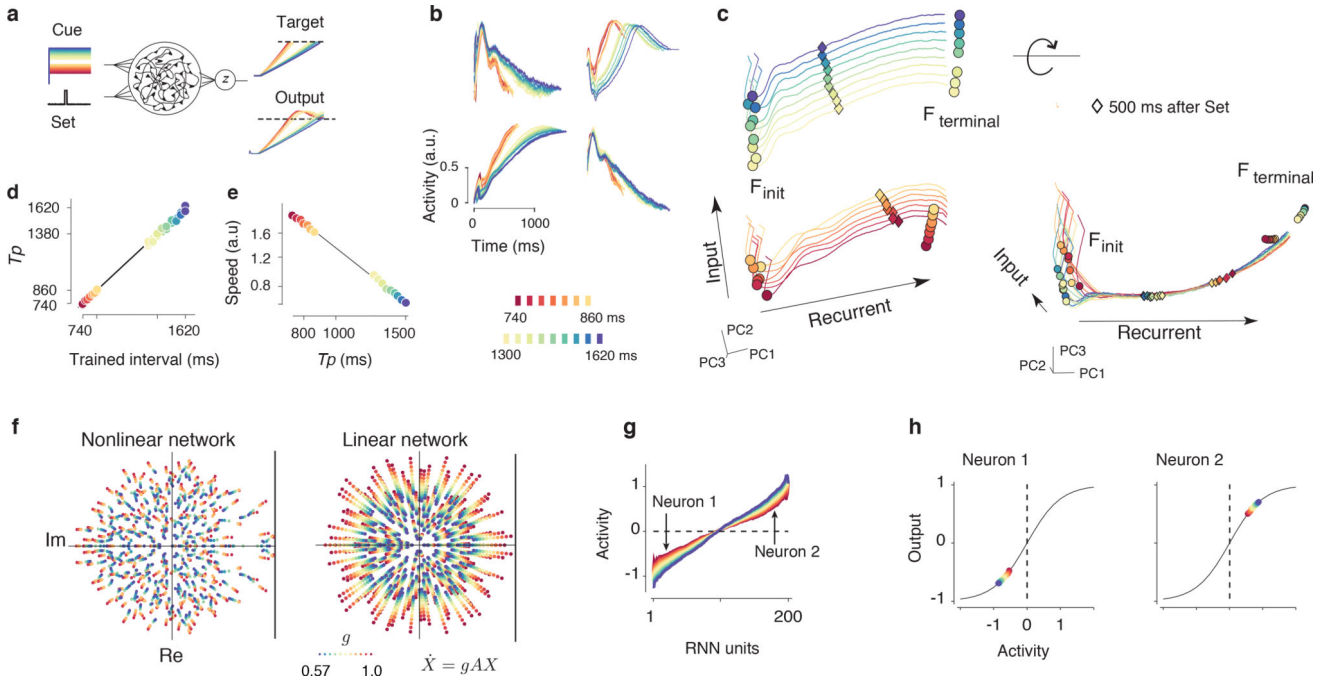

Fig. 5. Recurrent neural network model dynamics

(a) A recurrent neural network model that receives an input (Cue) whose strength depends on the desired interval (different colors), and a transient Set pulse that initiates the timing interval. The model produces a "response" when its output (z) reaches a fixed threshold. Network was trained to produce a linear ramp at its output. For other objectives see Supplementary Fig. 6. **(b)** The response profiles of randomly selected units aligned to the time of Set. Many units exhibit temporal scaling. **(c)** Left: Network activity projected onto the first three principal components (PCs) across all trials. Different traces correspond to trials with different durations (red for shortest to blue for longest). For each Cue input, the network engenders an initial and a terminal fixed point (circles; F_{init} and $F_{terminal}$). Diamonds mark the state of the network along the trajectory 500 ms after Set. The Cue input moves the fixed points within an "Input" subspace. The corresponding trajectories for different intervals reside in a separate "Recurrent" subspace. Right: Rotation of the state space reveals the invariance of trajectories in the recurrent subspace. In the recurrent subspace trajectories traverse the same path at different speeds (see diamonds for different Cue inputs). **(d)** After training, the network accurately produced the intervals according to the presented Cue input. **(e)** A plot of the average speed in the recurrent neural network model as a function of the production interval (*Tp*) on a log-log scale. The speed was estimated from the rate of change of activity along the neural trajectory within the subspace spanned by the first three PCs. **(f)** Left: The spectrum of eigenvalues of the linearized dynamics near $F_{terminal}$. Right: The spectrum of eigenvalues of an N-dimensional linear dynamical system $\tau\dot{x} = gAx$, with elements of A sampled from a Normal distribution $\mathcal{N}(0,1/N)$. Decreasing the gain values from $g = 1.0$ (red) to 0.57 (blue) progressively decreases the magnitude of the eigenvalues and increases the effective time constants $\tau_{eff} = \tau/g$. **(g)** Units in recurrent model were sorted based on their maximal activity when the network was near $F_{terminal}$. The plot shows the maximum activity as function of Cue input.

Vertical arrows mark two neurons, one with positive and another with negative activity, which are plotted in panel (h). **(h)** Stronger input drives units toward the saturation point of their nonlinear activation function where the shallowness of slopes leads to reduced gain of neural activity. This is true both for units with a positive response whose responses increased with Cue input (right), as well as units with a negative response, whose responses decreased with input drive (left). In all plots, different colors correspond to different intervals as shown by the color bar.

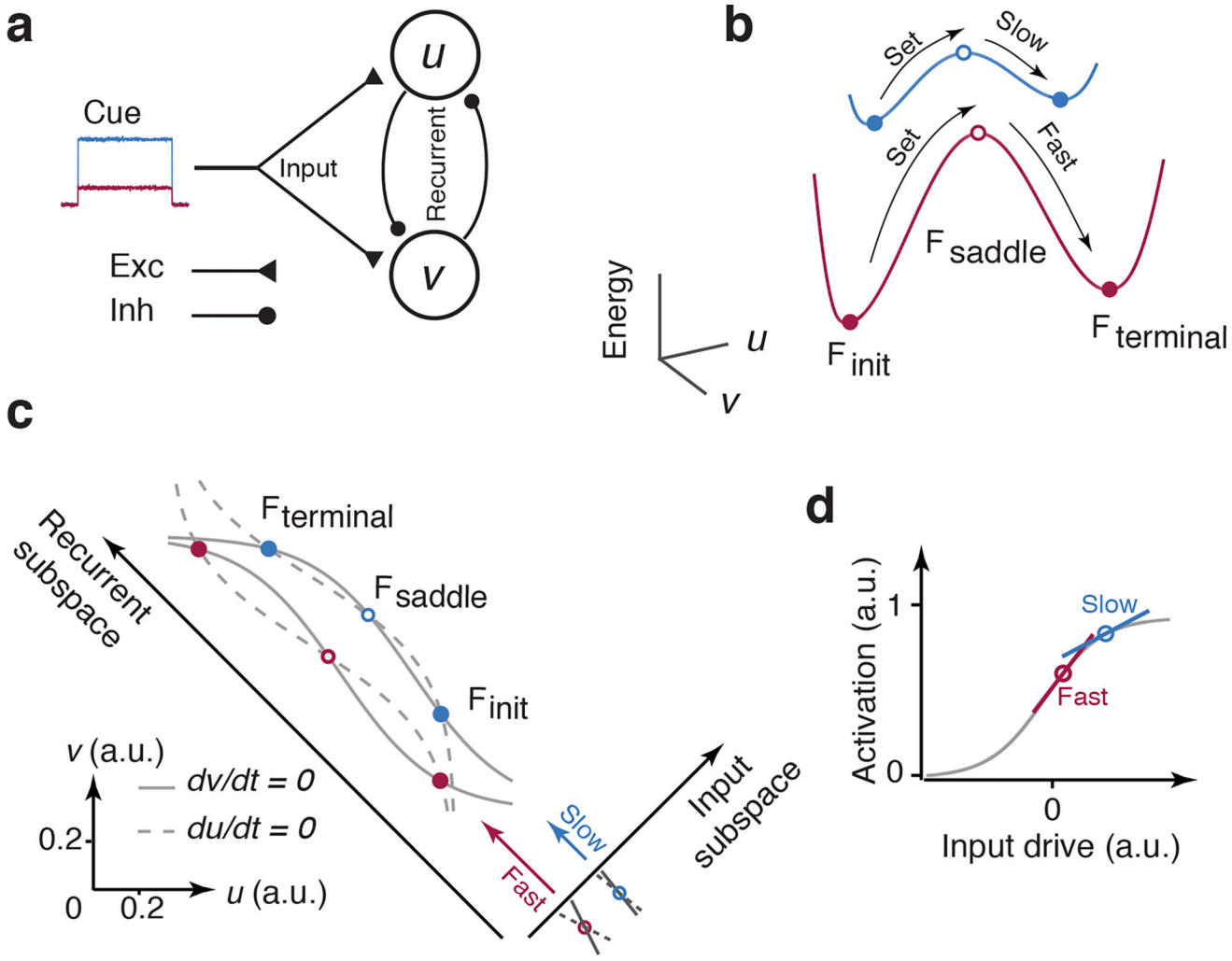

Figure 6. A simple two-neuron implementation of speed control

(a) Two inhibitory units (u and v) with recurrent inhibition receive a common excitatory input (Cue). **(b)** The energy landscape of the two-neuron model. The network has a bistable energy landscape whose gradients depend on the strength of the Cue input. Stronger inputs (blue) lead to shallower energy gradients and vice versa (red). The Set pulse moves the state away from the initial fixed point (F_{init}, filled circle) and over the saddle point (F_{saddle}, open circle). The network then spontaneously moves toward the terminal fixed point ($F_{terminal}$, filled circle). The speed of the movement toward $F_{terminal}$ is relatively slow when the energy gradient is shallow (blue) due to stronger common input. **(c)** Phase plane analysis of the 2-neuron model. The two axes on the lower left correspond to the activity of the two neurons (u and v). The input is applied to both units and thus drives the system along the diagonal, labeled as “input subspace”. The input level moves the sigmoidal nullclines of the two units (du/dt = 0, dashed, and dv/dt = 0, solid, see Supplementary Note) and adjusts the location of the three fixed points (F_{init}, $F_{terminal}$ and the intermediate F_{saddle}). The figure shows the two nullclines and the corresponding fixed points for two inputs levels (red and blue). Activation of Set moves the system along a “recurrent subspace” which is orthogonal to the input subspace. The proximity of nullclines (crosses below the Input subspace) controls the speed.

When the input is stronger, the nullclines are closer, which causes the system to become slower. **(d)** Interaction of the input drive with the saturating nonlinearity of one unit. The action of the input upon the nonlinear activation functions moves the saddle point and controls the speed of the system. Stronger inputs push the neurons toward the shallower part of the nonlinear activation function, and moves the saddle point to slower regions of the phase plane causing recurrent interactions to slow down.

Table 1

Effects of muscimol inactivation in the three brain areas

For the first four columns, we applied one-tailed paired-sample Student's *t*-tests to evaluate treatment effects on mean-squared-error in the main task. For the last column, we used two-tailed Student's *t*-tests to evaluate changes in reaction time in a control memory-guided saccade task. Each cell in the table reports the number of mini-sessions (n), the value of t-test, its degrees of freedom (t(df)), and the corresponding p values (*P*).

	Muscimol		**Saline**		
	Long	***Short***	***Long***	***Short***	***Reaction times in Muscimol***
MFC	n = 58, t(57) = 2.48, *P* = .002	n = 63, t(62) = 3.8, *P* = .003	n = 14, t(13) =1.39, *P* = .095	n = 14, t(13) = 1.13, *P* = .86	n = 12, t(11) = .066, *P* = .51
Caudate	n = 48, t(47) = 2.84, *P* = .005	n = 46, t(45) = 2.31, *P* = .036	n = 15, t(14) = 2.36, *P* = .60	n = 16, t(15) = 1.51, *P* = .86	n = 14, t(13) = 1.54, *P* = .26
Thalamus	n = 53, t(52) = 11.7, *P* << .001	n = 50, t(49) = 12.4, *P* << .001	n = 14, t(13) = 0.19, *P* = .81	n = 13, t(12) = 1.35, *P* = .12	n = 12, t(11) = 2.68, *P* = .0065

Table 2

Number of neurons recorded in each area

	MFC (included/total)	Caudate (included/total)	Thalamus (included/total)
Monkey A	281/356	101/200	481/534
Monkey D	135/166	177/309	365/402
Both animals	416/522	278/509	846/936