

Flexible timing by temporal scaling of cortical responses

Jing Wang^{1,5}, Devika Narain^{1,2,3,4}, Eghbal A. Hosseini^{2,5} and Mehrdad Jazayeri^{1,2,5*}

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 they 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 from 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 that the 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.

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 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 signaling properties of neurons across three brain areas that have been strongly implicated in timing: (i) the medial frontal cortex (MFC), which has been implicated in the inhibition¹, initiation^{2,3}, and coordination^{4–7} of movements, (ii) the caudate nucleus downstream of MFC, which is thought to play a major role in timing tasks^{8–15}, and (iii) thalamic regions that project to MFC and causally influence self-initiated movements¹⁶.

Neurons exhibited a diversity of complex response profiles that could not be reconciled with dominant models of timing¹³, including clock-accumulator models^{17,18}, oscillation-based models¹⁹, and 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. Notably, 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 mod-

els capable of using graded input to produce different time intervals. Analysis of these models revealed a previously uncharacterized 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 Long (1,500-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 (T_p) 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 the Short condition (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, $n=31$, $t_{30}=1.80$, $P=0.041$, and for monkey D, $n=35$, $t_{34}=6.44$, $P<0.001$).

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 distributions of within-session increases in the mean-squared error after the muscimol injection versus before the injection (for statistics, see Table 1). The drop in performance was due to a combination of changes in both bias and standard deviation (Fig. 2b). 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.

¹McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ³Netherlands Institute for Neuroscience, Amsterdam, The Netherlands. ⁴Department of Neuroscience, Erasmus Medical Center, Rotterdam, The Netherlands. Present address: ⁵Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. Jing Wang and Devika Narain contributed equally to this research. *e-mail: mjaz@mit.edu

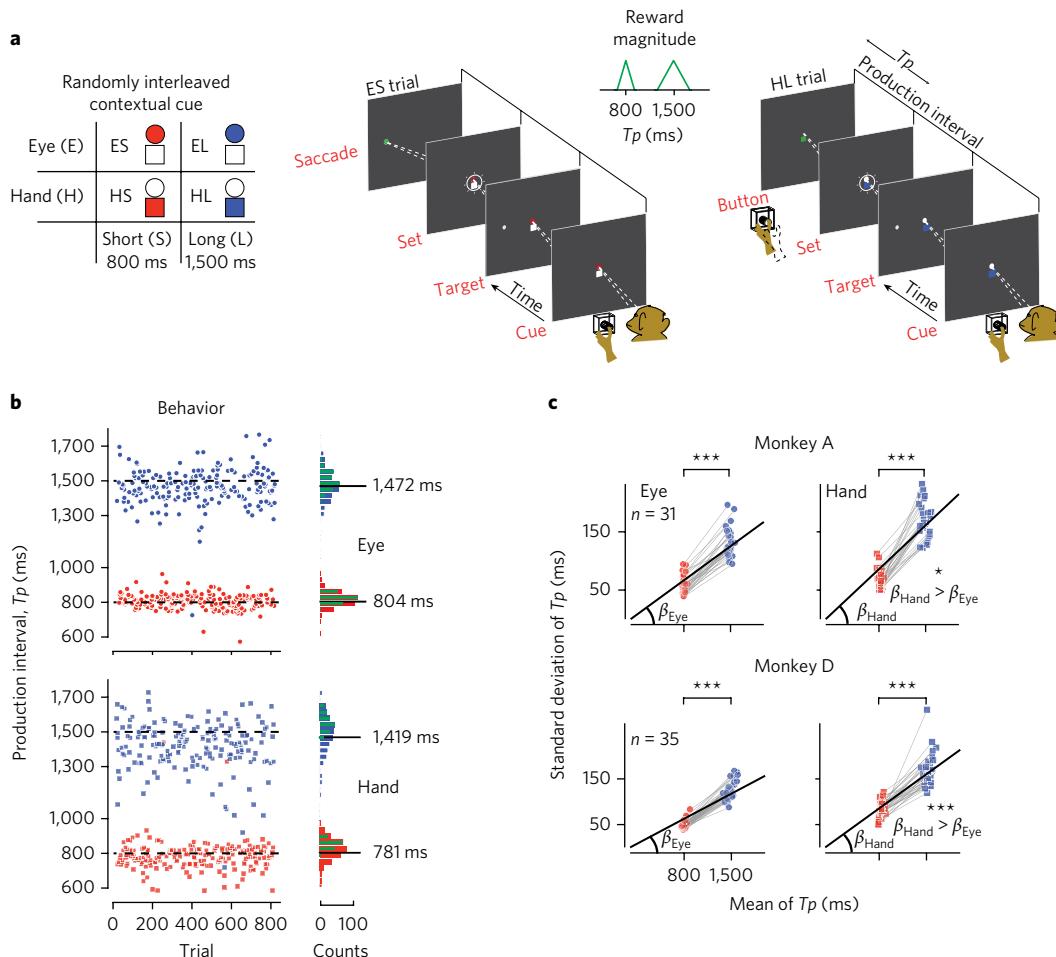


Fig. 1 | Time production task and behavior. a, Trial structure. Animals produced either an 800-ms (Short) or a 1,500-ms (Long) time interval, either by making a saccade (Eye) or a button press (Hand). These four conditions were randomly interleaved and were cued throughout the trial by the color and shape of two central stimuli, a circular fixation spot for Eye or a square fixation spot 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, Short; blue, 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 (T_p) was measured as the interval between Set and when either the saccade was made or the button was pressed. When T_p was generated with the desired effector and was within a specified reward window, the peripheral target (or square fixation) turned green, auditory feedback was provided, and the animal received a juice reward. 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% for monkey A and 51% for monkey D). The reward magnitude increased linearly with accuracy as shown by the green triangular reward function. Two example trials are shown, one for the Eye + Short (ES) condition (left) and one for the Hand + Long (HL) condition (right). **b**, A typical behavioral session showing T_p while the animal flexibly switched between the four trial conditions. For clarity, the Eye (top) and Hand (bottom) trials are plotted separately, although during the task they were randomly interleaved. The histograms on the right show the distribution of T_p for each condition with rewarded trials in green. Horizontal lines correspond to mean values, which are also reported numerically. **c**, For both effectors (left, Eye; right, Hand) and both animals (top, monkey A; bottom, monkey D), the s.d. of T_p scaled with mean T_p (red, Short; blue, Long). For monkey A, the mean \pm s.e.m. of T_p values across the conditions were ES: 810 ± 48.9 ms, Eye + Long (EL): $1,495 \pm 117$ ms, Hand + Short (HS): 822.3 ± 53.7 ms, HL: $1,486 \pm 136$ ms. For monkey D, they were ES: 808 ± 56.1 ms, EL: $1,481 \pm 137$ ms, HS: 836.7 ± 91.3 ms, HL: $1,521 \pm 177$ ms. The variability was significantly higher for the Long compared to the Short. The average Weber fraction (ratio of s.d. to mean) for the Hand (β_{Hand}) was significantly larger than Eye (β_{Eye} ; one-tailed paired-sample t test, for monkey A, $n=31$, $t_{30}=1.80$, $P=0.041$, and for monkey D, $n=35$, $t_{34}=6.44$, $***P<0.001$).

Temporal scaling of complex response profiles. To estimate each neuron's firing rate, we binned trials based on T_p 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 multimodal 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 polynomial fits of different degrees of freedom indicated that only 11% (47 of 416) of responses increased linearly; the rest were explained by higher-order polynomials. This number increased by only 4% when the

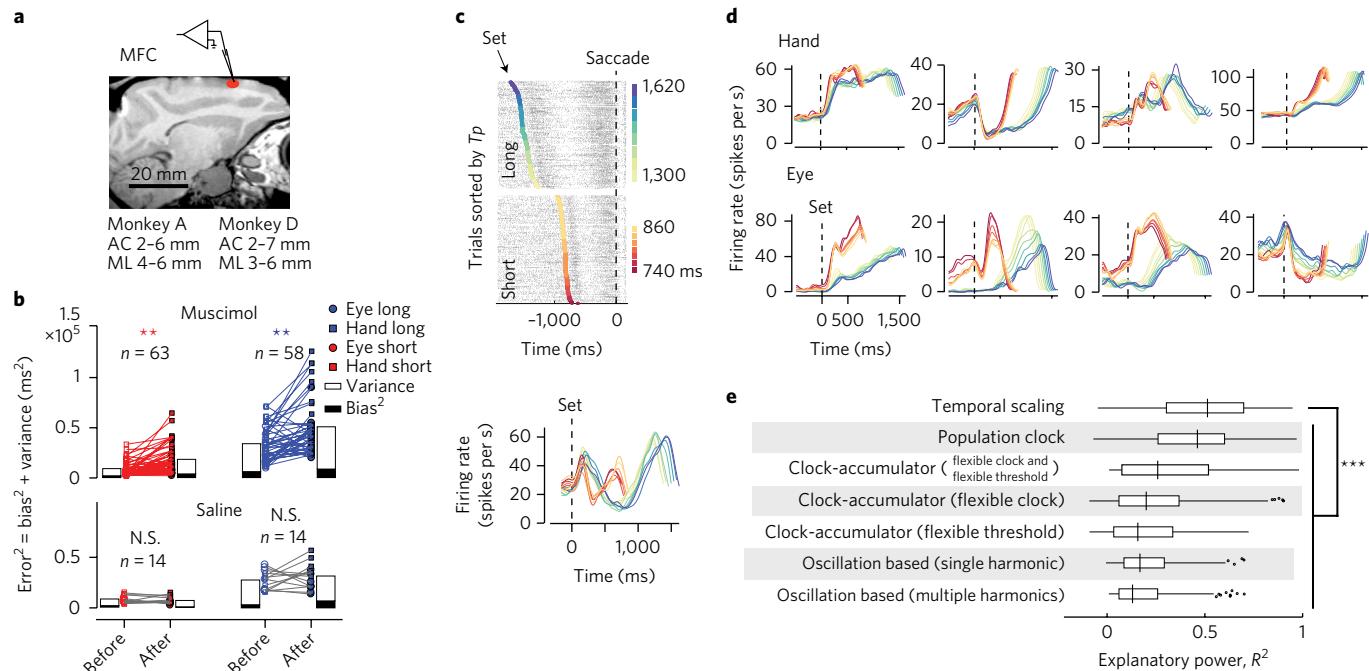


Fig. 2 | Medial frontal cortex inactivation and electrophysiology. **a**, Parasagittal view of the brain of one animal (monkey D) with a red ellipse showing the targeted region. Stereotaxic coordinates used in each animal are shown with respect to anterior commissure (AC) and midline (ML). **b**, Muscimol inactivation. Each line in each panel shows the change in mean squared error ($MSE = \sum(T_p - T_s)^2 = bias^2 + variance$; T_s is the desired interval) computed from minisession (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 bias² (black) and variance (white). Significance tests correspond to comparisons of MSE (see Table 1 for details) across minisessions (n , number of minisessions; ** $P < 0.001$; 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 (T_p). Bottom: average firing rates for each T_p 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 T_p bins. **d**, Activity profiles 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 plots 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, $\pm 1.5 \times$ 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, $F_{6,2859} = 125.2$, $P < 0.001$; one-tailed paired-sample t test between temporal scaling and population-clock model, $n = 416$, $t_{415} = 6.32$, *** $P < 0.001$). Models were cross-validated.

starting and terminating points of the linear ramps were allowed 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 and different frequencies¹⁹. In one variant, a single sinusoid was fit to the response of each neuron, and in another, multiple sinusoids (up to four) 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 response profile of each individual neuron is unique and context-independent, and the collective activity of the population determines 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 in the population clock model, the 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

for both random fluctuations of T_p within each temporal context (i.e., 800 ms or 1,500 ms) and deliberate adjustments of T_p across the two 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, $F_{6,2859} = 125.2$, $P < 0.001$).

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 T_p bins. This analysis revealed a wide range of SI values across the population (Supplementary Fig. 1a). 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 nonscaling signals (Fig. 3a), which was also evident from the distribution of SI values across individual neurons (Supplementary Fig. 1a).

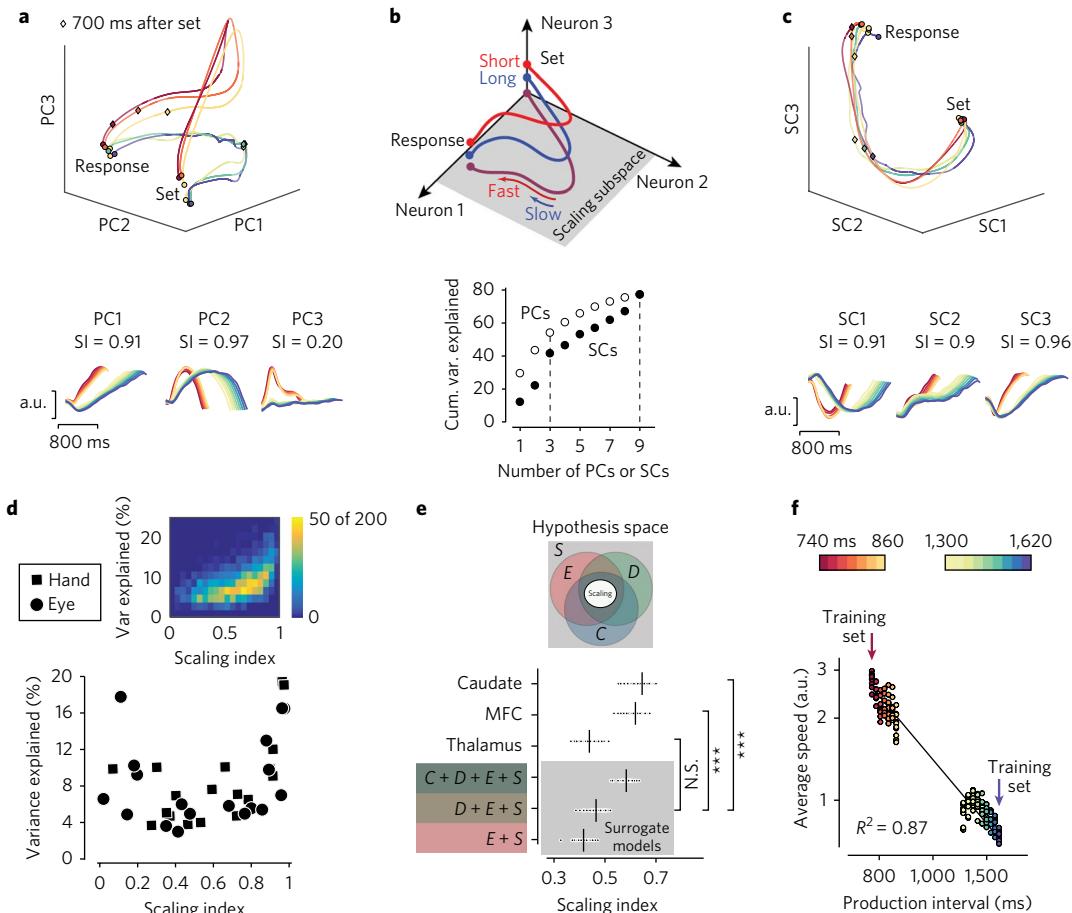


Fig. 3 | Temporal scaling in the medial frontal cortex at the population level. **a**, Top: population activity for Hand trials for Monkey A projected onto the first three PCs from the time of Set to the time of button press (Response). Activity profiles associated with different produced intervals are plotted in different colors (see color bar in **f**). Diamond shows activity 700 ms after Set. Bottom: time course of the first three PCs with the corresponding SI values. **b**, Top: schematic drawing illustrating the scaling subspace. The response dynamics associated with Short (red) and Long (blue) 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 (cum. var.) explained by PCs and SCs. **c**, Top: population activity sorted according to produced interval (T_p) bins and projected onto the first three SCs. As expected, in this subspace, the trajectories overlap. Bottom: the first three SCs with their corresponding SI values. Because of cross-validation, SIs were not in decreasing order (see text). **d**, Variance (var) 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 show 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). Inset: 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 in the MFC data, and the number of bootstrapped samples for each model was $n=200$. The plot shows the average SI across all SCs computed from bootstraps (small circles), along with the corresponding means (vertical lines) 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 thalamus (N.S., not significant; *** $P < 0.001$). **f**, The speed of neural trajectory within the scaling subspace spanned by the first three SCs predicted average T_p values across bins. The relationship between speed and T_p was fit to a linear log-log function. The scaling subspace was computed from training data (arrows, two T_p bins) and used to evaluate speed on the remaining test data (14 T_p bins). R^2 was computed by repeating the procedure using bootstrapping ($n=10$). Both axes are in log scale.

We hypothesized that perfect scaling might be found within a subspace of the population activity, i.e., a 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. 3a) had scaling indices of 0.91 and 0.97, respectively (Fig. 3a). 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 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 and 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 (data not shown). Responses projected

Table 1 | Effects of muscimol inactivation in the three brain areas

	Muscimol		Saline		
	Long	Short	Long	Short	Reaction times in muscimol
MFC	<i>n</i> =58	<i>n</i> =63	<i>n</i> =14	<i>n</i> =14	<i>n</i> =12
	<i>t</i> ₅₇ =2.48	<i>t</i> ₆₂ =3.8	<i>t</i> ₁₃ =1.39	<i>t</i> ₁₃ =1.13	<i>t</i> ₁₁ =0.066
	<i>P</i> =0.002	<i>P</i> =0.003	<i>P</i> =0.095	<i>P</i> =0.86	<i>P</i> =0.51
Caudate	<i>n</i> =48	<i>n</i> =46	<i>n</i> =15	<i>n</i> =16	<i>n</i> =14
	<i>t</i> ₄₇ =2.84	<i>t</i> ₄₅ =2.31	<i>t</i> ₁₄ =2.36	<i>t</i> ₁₅ =1.51	<i>t</i> ₁₃ =1.54
	<i>P</i> =0.005	<i>P</i> =0.036	<i>P</i> =0.60	<i>P</i> =0.86	<i>P</i> =0.26
Thalamus	<i>n</i> =53	<i>n</i> =50	<i>n</i> =14	<i>n</i> =13	<i>n</i> =12
	<i>t</i> ₅₂ =11.7	<i>t</i> ₄₉ =12.4	<i>t</i> ₁₃ =0.19	<i>t</i> ₁₂ =1.35	<i>t</i> ₁₁ =2.68
	<i>P</i> <0.001	<i>P</i> <0.001	<i>P</i> =0.81	<i>P</i> =0.12	<i>P</i> =0.0065

For the first four columns, we applied one-tailed paired-sample Student's *t* tests to evaluate treatment effects on MSE 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 minisessions (*n*), the value of the *t* test with its degrees of freedom (*t*_{df}), and the corresponding *P* values (*P*).

onto the subspace spanned by the first three SCs traced nearly identical trajectories that evolved at different speeds (Fig. 3c), which is precisely what is expected in the scaling subspace.

Next, we asked how much variance in the neural data the scaling subspace could 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 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). 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 and terminal firing rates, dimensionality, and the correlation between Short and Long activity profiles, but it 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 *Tps* were associated with slower speeds (Fig. 3f and Supplementary Fig. 4) and that 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. Notably, 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 and 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 that in 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 and 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 and see Methods). Consistent with previous work¹⁶, reversible inactivation strongly influenced animals' timing behavior (Table 1). However, several observations indicated that the function of thalamocortical signals was different from the functions of caudate and MFC signals (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*_{caudate}=278, Mann–Whitney–Wilcoxon test, *W*_{1,260}=310,733, *z*=7.89, *P*<0.001 for MFC; *W*_{1,120}=189,163, *z*=6.98, *P*<0.001 for caudate; Supplementary Fig. 1a). Second, scaling in the thalamus was significantly smaller than the *C+D+E+S* surrogate data that matched the neural data in terms of smoothness (S), endpoints (E), dimensionality (D), and correlation (C; one-tailed two-sample *t* test, *n*=200, *t*₃₉₈=35.2, *P*<0.001; 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 in 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 to scaling. 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 implies 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

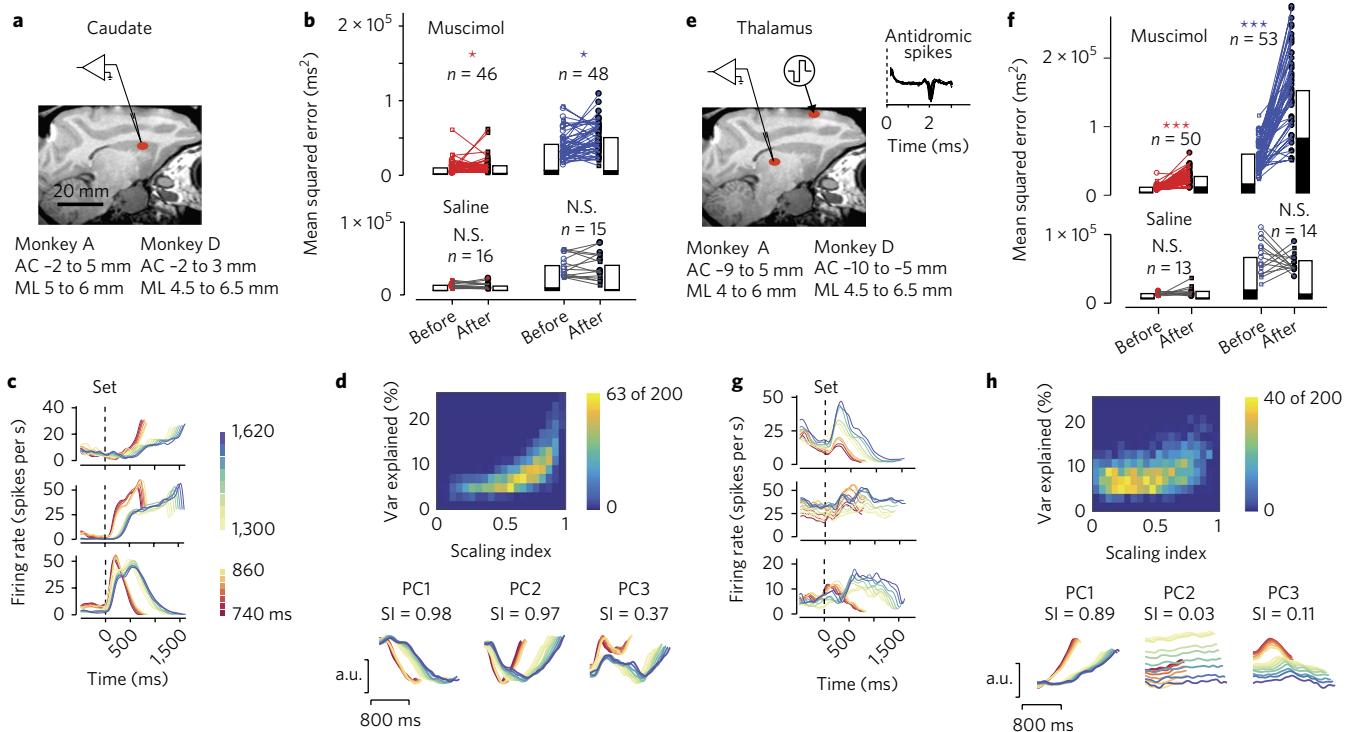


Fig. 4 | Inactivation, electrophysiology, and temporal scaling in the caudate and thalamus. **a**, As in Fig. 2a but with a red ellipse and stereotaxic coordinates showing targeted regions in the caudate. **b**, Muscimol inactivation in the caudate. Results are presented in the same format as in Fig. 2b. **c**, Activity profile of three example caudate neurons (same format as in Fig. 2d). **d**, Top: the relationship between variance (var) explained and SI in the caudate (same format as the inset in Fig. 3d). Bottom: the first three PCs with their corresponding SI values. **e**, As in **a** but 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 (format as in **b-d**). Responses in the thalamus are qualitatively different from the caudate (**d**) and MFC (Fig. 3d) in that most projections in the state space do not exhibit temporal scaling. N.S., not significant; * $P < 0.05$; *** $P < 0.001$.

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 the desired time³⁶.

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 nonscaling (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 scaling³⁷.

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 initial 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 changes 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 labor between these subspaces provides a simple explanation of why scaling and nonscaling signals might coexist within the same network. Nonscaling 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 that 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 that corresponds to the recurrent subspace in the model.

Notably, the model allows us to infer that within the nonscaling input subspace, production times should be correlated with the average level—not speed—of neural activity. To test this prediction, we investigated whether T_p could be predicted by the non-scaling component of MFC activity. We inferred the least-scaling

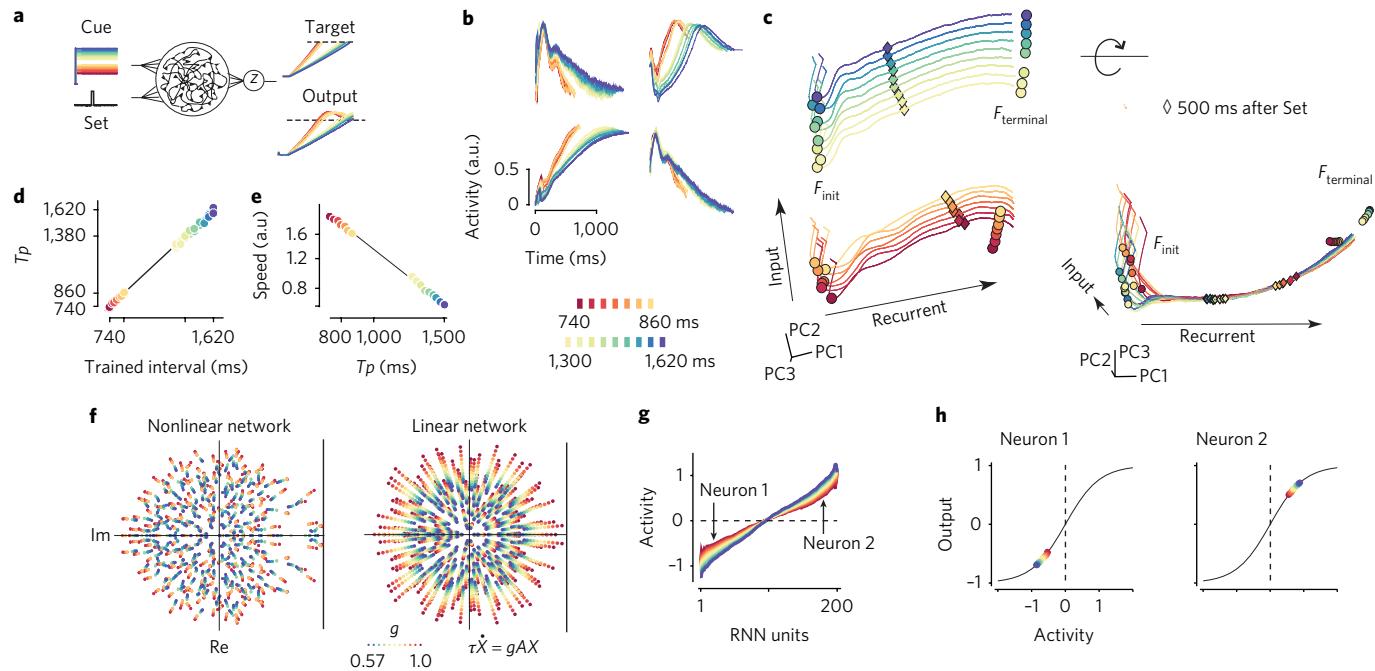


Fig. 5 | Recurrent neural network model dynamics. **a**, A recurrent neural network (RNN) model that receives an input (Cue) whose strength depends on the desired interval (different colors) and a transient Set pulse that initiates the timing interval. The model produces a ‘response’ when its output (z) reaches a fixed threshold (target). Network was trained to produce a linear ramp at its output. For other objectives, see Supplementary Fig. 6. **b**, 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 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**, Plot of the average speed in the recurrent neural network model as a function of the production interval (T_p) on a log-log scale. 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} (Re, real; Im, imaginary). Right: the spectrum of eigenvalues of an N -dimensional linear dynamical system $\tau \dot{X} = gAX$ (see Methods), with elements of A sampled from a normal distribution $\mathcal{N}(0, 1N)$. Decreasing the gain values from $g=1.0$ (red) to $g=0.57$ (blue) progressively decreases the magnitude of the eigenvalues and increases the effective time constants $\tau_{\text{eff}} = \tau/g$. **g**, Units in the RNN model were sorted based on their maximal activity when the network was near F_{terminal} . The plot shows maximum activity as a function of Cue input. Vertical arrows mark two neurons, one with positive and one with negative activity, plotted in **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) and for 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.

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 T_p to average MFC activity projected onto SC9. As predicted by the model, the average activities of the nonscaling components of MFC were indeed predictive of T_p (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 mechanism 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). 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). 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 constants³⁸. 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, similarly 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 in which 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

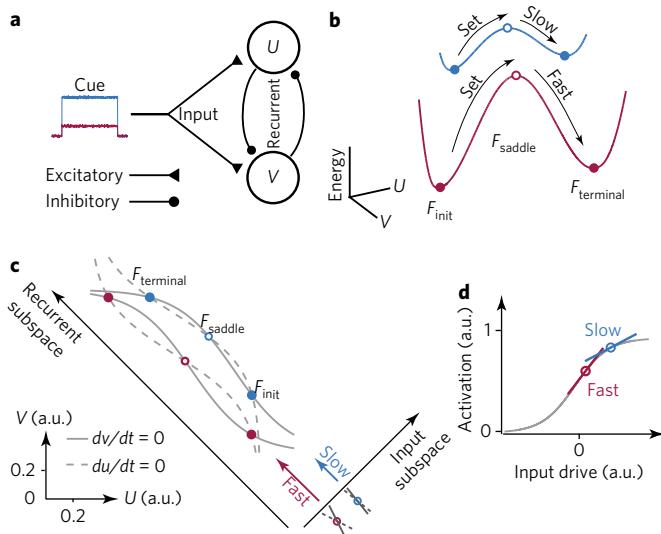


Fig. 6 | A simple two-neuron implementation of speed control. **a**, Two inhibitory units (U and V) with recurrent inhibition receive a common excitatory input (Cue). **b**, The energy landscape of the two-neuron model. The network has a bistable energy landscape whose gradients depend on the strength 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 two-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 (input subspace). The input level moves the sigmoidal nullclines of the two units ($du/dt = 0$, dashed; $dv/dt = 0$, 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 input 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 move the saddle point to slower regions of the phase plane, causing recurrent interactions to slow down.

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 points of units as a function of the input drive near F_{terminal} . Notably, 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 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 tradeoffs³⁹. Indeed, if the speed command is controlled by a sensory input, our recurrent network would behave similarly to more detailed network models consisting of excitatory and inhibitory units that approximate temporal integration of sensory information⁴⁰. 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 highlights 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 suggest⁴¹.

The source of the external input that adjusts the speed remains a pertinent and unresolved question. One possibility is that MFC receives this input directly from neurons in other cortical areas, which is consistent with recent observations in the parietal cortex⁴². 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 dynamics⁴⁵.

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 the 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 signaling 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.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at <https://doi.org/10.1038/s41593-017-0028-6>.

Received: 25 June 2017; Accepted: 29 October 2017;

Published online: 4 December 2017

References

1. Stuphorn, V. & Schall, J. D. Executive control of countermanding saccades by the supplementary eye field. *Nat. Neurosci.* **9**, 925–931 (2006).
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.* **36**, 3258–3268 (2012).
3. Fried, I. et al. Functional organization of human supplementary motor cortex studied by electrical stimulation. *J. Neurosci.* **11**, 3656–3666 (1991).
4. Lewis, P. A., Wing, A. M., Pope, P. A., Praamstra, P. & Miall, R. C. Brain activity correlates differentially with increasing temporal complexity of rhythms during initialisation, synchronisation, and continuation phases of paced finger tapping. *Neuropsychologia* **42**, 1301–1312 (2004).
5. Shima, K. & Tanji, J. Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *J. Neurophysiol.* **84**, 2148–2160 (2000).
6. Isoda, M. & Hikosaka, O. Switching from automatic to controlled action by monkey medial frontal cortex. *Nat. Neurosci.* **10**, 240–248 (2007).
7. Lu, X., Matsuzawa, M. & Hikosaka, O. A neural correlate of oculomotor sequences in supplementary eye field. *Neuron* **34**, 317–325 (2002).
8. Mello, G. B. M., Soares, S. & Paton, J. J. A scalable population code for time in the striatum. *Curr. Biol.* **25**, 1113–1122 (2015).
9. Gouvéa, T. S. et al. Striatal dynamics explain duration judgments. *eLife* **4**, e11386 (2015).
10. Jin, D. Z., Fujii, N. & Graybiel, A. M. Neural representation of time in cortico-basal ganglia circuits. *Proc. Natl. Acad. Sci. USA* **106**, 19156–19161 (2009).
11. Matell, M. S., Meck, W. H. & Nicolelis, M. A. L. Interval timing and the encoding of signal duration by ensembles of cortical and striatal neurons. *Behav. Neurosci.* **117**, 760–773 (2003).
12. Merchant, H., Harrington, D. L. & Meck, W. H. Neural basis of the perception and estimation of time. *Annu. Rev. Neurosci.* **36**, 313–336 (2013).
13. Mauk, M. D. & Buonomano, D. V. The neural basis of temporal processing. *Annu. Rev. Neurosci.* **27**, 307–340 (2004).
14. Bartolo, R., Prado, L. & Merchant, H. Information processing in the primate basal ganglia during sensory-guided and internally driven rhythmic tapping. *J. Neurosci.* **34**, 3910–3923 (2014).
15. Schultz, W. & Romo, R. Neuronal activity in the monkey striatum during the initiation of movements. *Exp. Brain Res.* **71**, 431–436 (1988).
16. Tanaka, M. Inactivation of the central thalamus delays self-timed saccades. *Nat. Neurosci.* **9**, 20–22 (2006).
17. Treisman, M. Temporal discrimination and the indifference interval. Implications for a model of the “internal clock”. *Psychol. Monogr.* **77**, 1–31 (1963).
18. Killeen, P. R. & Fetterman, J. G. A behavioral theory of timing. *Psychol. Rev.* **95**, 274–295 (1988).
19. Matell, M. S. & Meck, W. H. Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Brain Res. Cogn. Brain Res.* **21**, 139–170 (2004).
20. Karmarkar, U. R. & Buonomano, D. V. Timing in the absence of clocks: encoding time in neural network states. *Neuron* **53**, 427–438 (2007).
21. Buonomano, D. V. & Laje, R. Population clocks: motor timing with neural dynamics. *Trends Cogn. Sci.* **14**, 520–527 (2010).
22. Brody, C. D., Hernández, A., Zainos, A. & Romo, R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb. Cortex* **13**, 1196–1207 (2003).
23. Komura, Y. et al. Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature* **412**, 546–549 (2001).
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. USA* **108**, 19784–19789 (2011).
25. Emmons, E. B. et al. Rodent medial frontal control of temporal processing in the dorsomedial striatum. *J. Neurosci.* **37**, 8718–8733 (2017).
26. Gibbon, J. Scalar expectancy theory and Weber’s law in animal timing. *Psychol. Rev.* **84**, 279 (1977).
27. Rakitin, B. C. et al. Scalar expectancy theory and peak-interval timing in humans. *J. Exp. Psychol. Anim. Behav. Process.* **24**, 15–33 (1998).
28. Creelman, C. D. Human discrimination of auditory duration. *J. Acoust. Soc. Am.* **34**, 582–593 (1962).
29. Gibbon, J., Church, R. M. & Meck, W. H. Scalar timing in memory. *Ann. NY Acad. Sci.* **423**, 52–77 (1984).
30. Grondin, S. From physical time to the first and second moments of psychological time. *Psychol. Bull.* **127**, 22–44 (2001).
31. Hikosaka, O., Sakamoto, M. & Usui, S. Functional properties of monkey caudate neurons. I. Activities related to saccadic eye movements. *J. Neurophysiol.* **61**, 780–798 (1989).
32. Parthasarathy, H. B., Schall, J. D. & Graybiel, A. M. Distributed but convergent ordering of corticostriatal projections: analysis of the frontal eye field and the supplementary eye field in the macaque monkey. *J. Neurosci.* **12**, 4468–4488 (1992).
33. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
34. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532**, 459–464 (2016).
35. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
36. Hanes, D. P. & Schall, J. D. Neural control of voluntary movement initiation. *Science* **274**, 427–430 (1996).
37. Sussillo, D. & Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput.* **25**, 626–649 (2013).
38. Machens, C. K., Romo, R. & Brody, C. D. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science* **307**, 1121–1124 (2005).
39. Hanks, T., Kiani, R. & Shadlen, M. N. A neural mechanism of speed-accuracy tradeoff in macaque area LIP. *eLife* **3**, e02260 (2014).
40. Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955–968 (2002).
41. Meck, W. H. Neuropharmacology of timing and time perception. *Brain Res. Cogn. Brain Res.* **3**, 227–242 (1996).
42. Jazayeri, M. & Shadlen, M. N. A neural mechanism for sensing and reproducing a time interval. *Curr. Biol.* **25**, 2599–2609 (2015).
43. Buhusi, C. V. & Meck, W. H. Differential effects of methamphetamine and haloperidol on the control of an internal clock. *Behav. Neurosci.* **116**, 291–297 (2002).
44. Soares, S., Atallah, B. V. & Paton, J. J. Midbrain dopamine neurons control judgment of time. *Science* **354**, 1273–1277 (2016).
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* **88**, 419–431 (2015).

Acknowledgements

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

Author contributions

J.W. was responsible for all aspects of experiments and 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 interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41593-017-0028-6>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.J.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Methods. Two adult rhesus monkeys (*Macaca mulatta*, a 6.5-kg female and a 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 (<https://mworks.github.io/>) running on a Mac Pro was used to deliver stimuli and to control behavioral contingencies. Visual stimuli were presented on a 23-inch (58.4-cm) 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° 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 1,500 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 (1,500-ms) intervals. After a delay period (500–1,500 ms, uniform hazard), the saccade target was briefly presented 8° 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 (500–1,500 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 the 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 circular window of radius 2.5° 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 hemifield⁴⁶. 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-min warm-up 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 multiunit 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 time-stamped at 30 kHz and streamed to a data acquisition system (OpenEphys).

The dataset collected for this study included 1,967 single units or multiunits recorded from the MFC, caudate and thalamus of two monkeys (Table 2), in which 69% (1,351/1,967) were tentatively single units. Neurons with firing rates less than 2 spikes per s 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 30 G stainless steel injection cannula via a fused silica injection line (365-μm OD, 100-μm ID, Polymicro Technologies). In each injection session, we first established the animal's baseline behavioral performance. Afterwards, we pressure-injected muscimol hydrobromide (5 μ

g/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 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 numbers of trials in the before and after muscimol injection. For statistical comparison, these inequalities may introduce sampling biases. To avoid such biases, we created 50-trial minisessions from before and after the injections, in which the trials within a minisession 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, $MSE = \sum (T_p - T_s)^2 / Bias^2 + Var$, before and after the injection for every minisession. The same procedure was used to assess the 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 μA) 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 (\mathbf{x}), bold and uppercase for matrices (X). Brackets were used for indexing vectors and matrices ($\mathbf{x}[i]$ and $\mathbf{X}[i,j]$). Subscripts were used for indexing a set of scalars (x_i), vectors (\mathbf{x}_i), or matrices (X_i). Subscripts were also used to show projections onto a subspace. For example, $\mathbf{x}_{PC(1:k)}$ refers to a vector projected onto the first k principal components. Curly brackets were used to indicate a subset of conditions. For example, $\{\mathbf{x} | a = a_0; b = b_1\}$ 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 \cup was used to indicate data combined across a number of variables. For example $\bigcup_i^N \{\mathbf{x}_i\}$ denotes data collected across a union of vectors \mathbf{x}_i . The symbol $\langle \mathbf{x} \rangle_i$ was used to show averaging of a vector \mathbf{x} across i . Point functions were shown as lowercase ($f(\cdot)$) 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 and 7 kHz, and spike waveforms were detected at a threshold that was typically set to $3 \times$ the RMS noise. Single units and multiunits were sorted offline using 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 T_p values more than 3 s.d. away from the mean (1.46% of trials). Firing rates were estimated by (i) averaging spike counts per time bin, (ii) using a 40-ms Gaussian kernel to compute smooth spiking density functions, and (iii) z-scoring to minimize sampling bias due to baseline and amplitude differences across neurons.

To examine the relationship between firing rates and T_p values, we binned trials according to T_p and compared average firing rates for each bin. For the 800-ms interval, we used seven bins centered on 740 to 860 ms every 20 ms, and for the 1,500-ms interval, we used nine bins centered on 1,300 to 1,620 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 T_p bin by $r(t;c,T_p)$. For population analyses, response vectors of individual neurons were organized into rows of a matrix denoted by $r(t;c,T_p)$.

To test whether activity profiles could be described by a linear function (for example, ramping activity), we compared 0-order 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.

Comparing the motor timing models at the level of single- and multiunits. 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

Table 2 | Numbers 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

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 the 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 four 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 were used to find the best-fitting polynomial, and the temporally scaled versions 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 nine principal components (PCs) captured nearly 80% of the variance in the data (Fig. 3b). We therefore computed the scaling components (SCs) from data captured by the first nine PCs, which was computed as follows:

$\mathbf{r}_{PC}(t; c) = \widehat{\mathbf{V}}^* \mathbf{r}(t; c)$ where $\widehat{\mathbf{V}} = [v_1^T; v_2^T; \dots; v_{N_{PC}}^T]$ is the projection matrix and v_i is the i th PC direction. Therefore, the denoised activity across all conditions and time points $\mathbf{r}_{PC}(t; c)$ is of size $N_{PC} \times (T \times C)$. We computed the corresponding scaled responses using our scaling procedure and denoted the result by $\mathbf{r}_{SC}^S(t; c)$. To find the scaling subspace, we solved an optimization problem that minimized the difference between average firing rates associated with different T_p values (for example, T_p and T_p'). We denote the corresponding projection by U_{SC} and refer to its columns as scaling components (SCs). The resulting projection \mathbf{r}_{SC} can be computed as follows:

$$\mathbf{U}_{SC} = \arg \min_U \left\{ \text{var} \left[U^* \left[\mathbf{r}_{PC}^S(t; T_p) - \mathbf{r}_{PC}^S(t; T_p') \right] \right] \right\}$$

$$\mathbf{r}_{SC}(t; c) = \mathbf{U}_{SC}^* \mathbf{r}_{PC}(t; c)$$

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

$$S(T_p) = \frac{1}{T} \sum_{t=1:T} \| d\mathbf{r}_{SC1:3}(t, T_p) / dt \|$$

$$\mathbf{r}_{SC1:3}(t; c) = \mathbf{U}_{SC1:3}^* \mathbf{r}_{PC}(t; c)$$

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

$$\log(S(T_p)) = A - B \cdot \log(T_p)$$

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: (i) trials were sorted based on production interval (T_p); (ii) sorted trials were grouped into bins of similar T_p values (as described above); (iii) the first nine PCs and the corresponding SCs for each bin were computed; and (iv) 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 T_p values.

$$z_{scaled} = \mathbf{r}_{PC/SC}^S(t; \bigcup_i^N \{T_p_i\})$$

$$SI = \frac{\sum_{t=1}^{N_t} [z_{scaled} - \langle z_{scaled} \rangle_{T_p}]^2}{\sum_{t=1}^{N_t} \sum_{T_p=T_p}^{T_p'} [\langle z_{scaled} \rangle_{T_p} - \langle z_{scaled} \rangle_{T_p'}]^2}$$

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:

$$\dot{x}(t) = \tau x(t) - x + J r(t) + B u + c_x + \rho(t)$$

$$r(t) = \tanh(x(t))$$

Variable $x(t)$ is an N -dimensional vector representing the activity of all the units. Variable $r(t)$ represents the firing rates of those units by transforming x through a \tanh saturating nonlinearity. The time constant of each neuron was set to $\tau = 10$ ms. This value is different from τ_{eff} , which emerges at the network level. Variable c_x is a vector representing a stationary offset the units receive, and $\rho(t)$ is a vector representing white noise $N(0, 0.01)$ sampled at each time-step $\Delta t = 1$ ms. The recurrent connections in the network are specified by matrix J , whose initial 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 $B = [b_c, b_s]$, 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. The value of u_c was set to 0 for 100 ms and then jumped to a graded value proportional to the length of one of 16 desired intervals distributed within a range 500–1,700 ms. The offset of u_c was sampled proportionally from the range 0.1 to 0.6 and was perturbed with Gaussian noise $N(0, 0.25)$ 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 w_o and a bias term c_z . The output weights were initialized to zero at the start of training.

$$z(t) = w_o^T r(t) + c_z$$

Statistics. The Weber fractions across behavioral sessions (Fig. 1c), MSEs before and after inactivation (across minisessions; Table 1 and Figs. 2a and 4b,f), scaling indices obtained from a bootstrap procedure for various brain areas, and surrogate data (Fig. 3e) were assumed to be normally distributed, but 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 are reported in Table 2. For single-neuron responses with respect to the seven types of timing models, we used one-way ANOVA 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 pairs of brain areas under examination (thalamus and MFC, thalamus and caudate).

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

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

References

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.* **87**, 2158–2166 (2002).
47. Namboodiri, V. M. & Hussain Shuler, M. G. Report of interval timing or action? *Proc. Natl. Acad. Sci. USA* **111**, E2239–E2239 (2014).
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. USA* **111**, 480–485 (2013).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

► Experimental design

1. Sample size

Describe how sample size was determined.

Approximately 2000 neurons were recorded from 2 subjects, which is larger than most existing studies in the literature. No statistical methods were used to predetermine sample sizes. Details have been reported in Table 1 in methods.

2. Data exclusions

Describe any data exclusions.

Neurons with firing rates less than 2 spikes per second during the timing epoch were excluded from further analyses. According to this criterion for both animals 416 out of 522 neurons were included for the medial frontal cortex, 278 out of 509 for the Caudate and 846 of 936 neurons were included.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Both animals were trained over a period of 2 years and performed thousands of trials. Results were highly consistent (reported in Fig. 1) Causal experiments were performed with multiple sessions and thousands of trials. See Methods for details.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No experimental groups were assigned, therefore randomization is not relevant to this study.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No group allocation was performed during data collection or analysis, therefore blinding is not relevant.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

All analysis used MATLAB 2016b, Single- and multi-units were sorted using a custom MKsort (Version 2, <https://github.com/ripple-neuro/mksort>).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). [Nature Methods guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used in this study

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in this study

10. Eukaryotic cell lines

- State the source of each eukaryotic cell line used.
- Describe the method of cell line authentication used.
- Report whether the cell lines were tested for mycoplasma contamination.
- If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No Eukaryotic cell lines were used in this study

NA

NA

NA

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Two adult rhesus monkeys (Macaca mulatta, a 6.5 kg female and 9.0 kg male, both 5 years old)

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

This study did not involve human research participants