bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Reinforcement regulates timing variability in thalamus

Jing Wang[1,2], Eghbal Hosseini[2], Nicolas Meirhaeghe[3], Adam Akkad[2], Mehrdad Jazayeri[2,*]
[1]Department of Bioengineering, University of Missouri, Columbia, MO 65211 [2]McGovern Institute for Brain Research, Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 [3]Harvard-MIT Division of Health Sciences & Technology, Cambridge, Massachusetts 02139, USA

*Corresponding author

Mehrdad Jazayeri, Ph.D.
Robert A. Swanson Career Development Professor
Assistant Professor, Department of Brain and Cognitive Sciences
Investigator, McGovern Institute for Brain Research
Investigator, Center for Sensorimotor Neural Engineering
MIT 46-6041
43 Vassar Street
Cambridge, MA 02139, USA
Phone: 617-715-5418
Fax: 617-253-5659
Email: mjaz@mit.edu

Acknowledgements

M.J. is supported by NIH (NINDS-NS078127), the Sloan Foundation, the Klingenstein Foundation, the Simons Foundation, the McKnight Foundation, the Center for Sensorimotor Neural Engineering, and the McGovern Institute. N.M. is supported by the Center for Sensorimotor Neural Engineering.

Author contribution

J.W., E.H., and M.J. conceived the project. J.W. and E.H. collected the main behavioral and electrophysiology data. J.W. analyzed the data and developed the model. E.H. played a major role in data analysis. J.W and M.J. interpreted the data with contribution from E.H. and N.M.. N.M. performed and analyzed the control experiment in a third monkey, and highlighted the need for additional validation analyses. A.A. conducted the human psychophysics experiments. M.J. supervised the project. All authors contributed to the writing of the manuscript.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Motor learning reduces variability and motor variability facilitates exploratory learning. This paradoxical relationship has made it challenging to tease apart behavioral and neural signatures of variability that degrade performance from those that improve it. To tackle this question, we analyzed behavioral variability and its correlates across populations of neurons in the premotor cortex, thalamus and caudate within the cortico-basal ganglia circuits during a flexible motor timing task. Behavioral variability was comprised of two opposing processes: a slow drift in memory that caused effector and interval specific variability, and a fast reinforcement learning process that countered memory drifts in a context-specific manner. In all three brain regions, memory drifts were accompanied by context-specific fluctuations of neural activity along behaviorally-relevant dimensions. However, only in thalamus was drift-related neural variability regulated by reinforcement. Previously, we found that thalamus provides a speed command to control response dynamics in cortex and caudate responsible for flexible motor timing. Our current findings indicate that the nervous system uses reinforcement to regulate the variability of the speed command in thalamus in order to calibrate memory in the presence of inherent slow drifts. More generally, our work provides an example of reinforcement acting upon neural variability to improve behavioral performance.

Introduction

While interacting with a dynamic and uncertain environment, acting variably can sometimes be beneficial. A prime example of this is in the motor system where variability can facilitate learning (Dhawale et al., 2017). However, the relationship between variability and learning is nuanced. On the one hand, motor learning reduces variability (Crossman, 1959; Harris and Wolpert, 1998; Smith et al., 2006; Sternad and Abe, 2010; Thoroughman and Shadmehr, 2000; Verstynen and Sabes, 2011). This is evident in myriad behaviors such as a child learning to touch his nose or an olympian perfecting her most sophisticated move on an ice rink. On the other hand, motor variability may play a direct role in the induction of learning (Dhawale et al., 2017). When acquiring a new motor skill or adapting an old skill to a new environment, being variable enables learning through trial and error. Indeed, humans learn more efficiently if the structure of their natural movement variability aligns with the underlying learning objective (Wu et al., 2014).

One intriguing hypothesis is that the brain regulates motor variability to facilitate learning (Huang et al., 2011; Kao et al., 2005; Olveczky et al., 2005; Tumer and Brainard, 2007). This seems particularly relevant when the success or failure of a movement is determined by a binary or scalar (unsigned) feedback without any graded sensory error (Chen et al., 2017; Dam et al., 2013; Izawa and Shadmehr, 2011; Nikooyan and Ahmed, 2015; Pekny et al., 2015; Shmuelof et al., 2012; Wu et al., 2014). As formalized in the theory of reinforcement learning, when errors are binary and/or unsigned, the system has to explore the action space, probe and update the value of different possibilities and adjust the frequency with which those are expressed (Sutton and Barto, 1998). Several indirect lines of evidence support this hypothesis. For example,

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

reaching movements become more variable during periods of low success rate (Izawa and Shadmehr, 2011; Pekny et al., 2015), and metrics of saccadic eye movements become more variable in the absence of reward (Takikawa et al., 2002). These experiments highlight a potential reciprocal relationship between learning and variability where variability facilitates learning and learning reduces variability.

While statistical treatment of behavior has been consistent with reward-dependent adjustment of variability in certain behavioral domains, a rigorous assessment of this hypothesis demands two important developments. First, we need a model capable of teasing apart sources of variability that hinder performance from those that facilitate learning. Second, we need evidence that the underlying neural circuits rely on reinforcement to regulate their variability along task-relevant dimensions. To address these problems, we focused on a motor timing task in which monkeys had to produce different time intervals using different effectors. We first verified the presence of reward-dependent adjustment of variability in motor timing. We then developed a generative model to explain how reward regulates variability and facilitates learning. Finally, we probed the underlying neural circuits to ask whether reward could regulate task-relevant variability within the nervous system. We recorded from multiple nodes of the cortico-basal ganglia circuits that were previously found to support monkeys' timing behavior (Wang et al., 2018). Results indicated that the variability across the population of thalamic neurons with projections to the dorsomedial frontal cortex (DMFC) was indeed regulated by reward in a context-specific manner.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results

Cue-Set-Go task

Two monkeys were trained to perform a Cue-Set-Go (CSG) motor timing task (Figure 1A). On each trial, the animal aimed to produce either an 800 ms (Short) or a 1500 ms (Long) time interval (t_t) either with a saccade (Eye) or with a button press (Hand). Each trial was initiated by presenting a fixation spot (“Cue”) at the center of the screen. The Cue consisted of a circle and a square, and specified the trial type. For Eye trials, the square was white and the color of the circle indicated the interval: red for Short and blue for Long. For Hand trials, the circle was white and the interval was cued by the color of the square. The four trial types, Eye-Short (ES), Eye-Long (EL), Hand-Short (HS), and Hand-Long (HL) were randomly interleaved throughout the session. After a random delay, a visual stimulus (“Tar”) was flashed to the left or right of the screen. This stimulus specified the position of the saccadic target for the Eye trials and served no function in the Hand trials. After another random delay, the presentation of a “Set” flash around the fixation spot indicated the start of timing period. Animals had to proactively initiate a saccade or a button press aiming at the desired t_t. We refer to the movement initiation as the “Go”. We measured the produced interval, t_p, as the interval between Set and Go. To receive reward, the relative error, defined as $e = (t_p\text{-}t_t)/t_t$, had to be within a reward window (Figure 1A). On rewarded trials, the magnitude of reward decreased linearly with the size of the error. The width of the reward window was controlled independently for each trial type and was adjusted adaptively on a trial-by-trial basis (see Methods).

Animals learned to use the Cue and flexibly switched between the four trial types (Figure 1B). For both effectors, a robust feature of the behavior was that produced intervals (t_p) were more variable for the Long compared to Short (Figure 1C). This is consistent with the common observation that timing variability scales with the interval being timed (Gibbon, 1977; Malapani and Fairhurst, 2002). To quantify this *scalar variability* in our dataset, we measured the mean (μ) and standard deviation (σ) of t_p in each behavioral session separately for each trial type, and used linear regression ($\sigma(t_p) = \beta_1\, \mu(t_p)$) to test the relationship between σ and μ for each animal and each effector (Figure 1C). For both animals and effectors, the slope was significantly positive (one-tailed t-test, *** $p << 0.001$, for monkey A, df = 128, t = 32.12; for monkey D, $p << 0.001$, df = 163, t = 24.06).

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

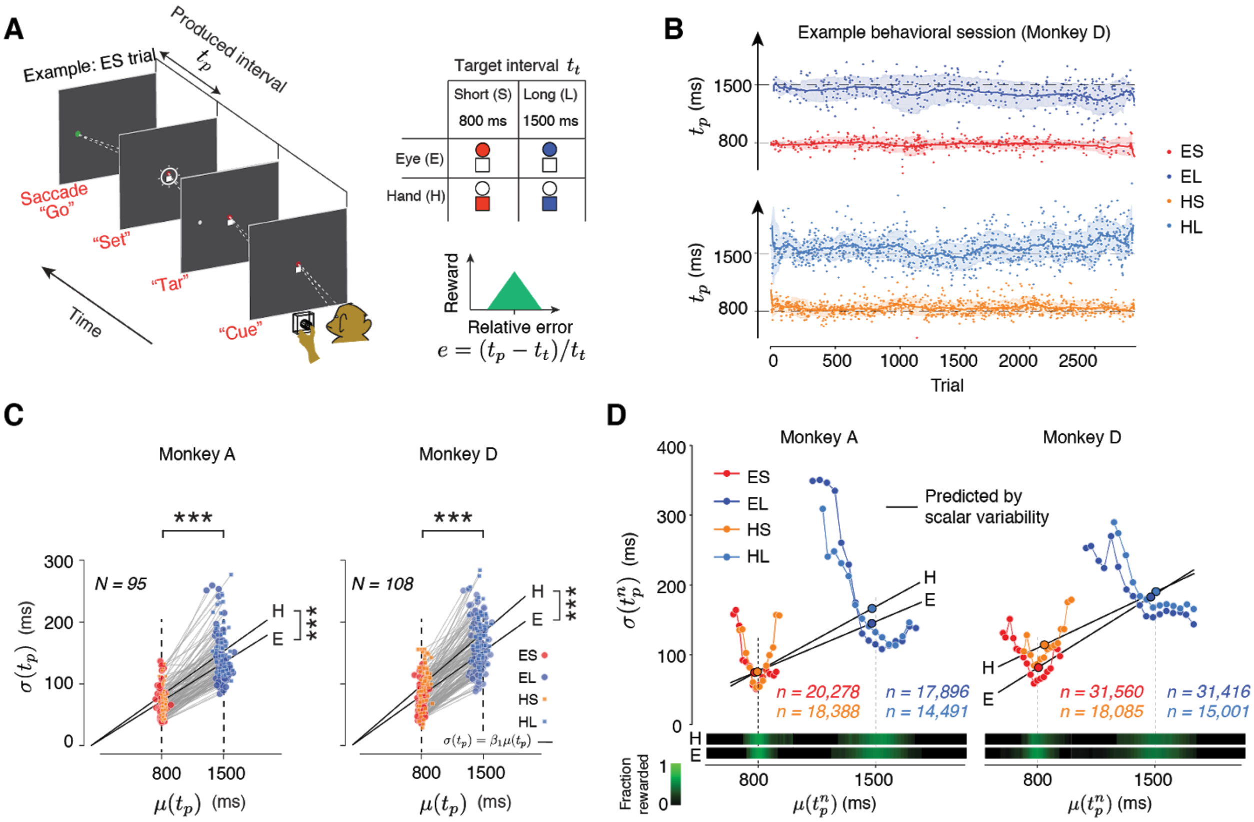

Figure 1. Task, behavior, and reward-dependency of variability. (A) The Cue-Set-Go task. The animal has to produce one of two target intervals, t_t, of 800 ms (Short) or 1500 ms (Long) either by a timed saccade (Eye) or a timed button press (Hand). Every trial begins when the animal fixates a central spot. The fixation spot, referred to as the "Cue", instructs the animal about the desired interval and effector (top right). After fixation, a stimulus is flashed to the left or right of the fixation ("Tar"). After another random delay, a white ring is flashed around the fixation spot ("Set"). The produced interval, t_p, is the interval between Set and the motor response ("Go"). The four trial types are randomly interleaved. The left panel shows an Eye-Short (ES) trial. Bottom right: The animal receives reward if it responds with the desired effector and if the relative timing error (e = (t_p-t_t)/t_t) is within the acceptance window (shaded green). The amount of reward drops linearly with the magnitude of relative error. (B) Animal's behavior in a representative session. For visual clarity, t_p values (dots) for different trial types are shown along different abscissae for each trial type. The solid line and shaded area are the mean and standard deviation of t_p calculated from a 50-trial sliding window. (C) Standard deviation of t_p ($\sigma(t_p)$) as a function of its mean ($\mu(t_p)$) for each trial type in each behavioral session. Each pair of connected dots corresponds to Short and Long of the same effector in a single session. In both animals, the variability was significantly larger for the Long compared to the Short for both effectors (one-tailed paired-sample t test, *** p <<0.001, for monkey A, n =190, t_{128} = 157.4; for monkey D, n = 216, t_{163} = 181.7). The solid black lines show the regression line relating $\sigma(t_p)$ to $\mu(t_p)$ across all behavioral sessions for each trial type ($\sigma(t_p)$ = β_1 $\mu(t_p)$). Regression slopes were positive and significantly different from zero for both effectors (β = 0.087 ± 0.02 mean ± std for Eye and 0.096 ± 0.021 for Hand in Monkey A; β = 0.10 ± 0.02 for Eye and 0.12 ± 0.021 for Hand in Monkey D). Hand trials were more variable than Eye ones (one-tailed paired-sample t-test, for monkey A, n = 95, t_{52} = 6.92, *** p <<0.001, and for monkey D, n = 108, t_{61} = 6.51, ***p << 0.001). (D) Dependence of timing variability on reward rate. Top: The plot shows the relationship between the standard deviation and mean of t_p ($\sigma(t_p)$ and $\mu(t_p)$) when these statistics are derived locally. To create this plot, we (1) computed local estimates of $\mu(t_p)$ and $\sigma(t_p)$ for trials of the same type using a 50-trial sliding window, (2) gathered all the local estimates across all behavioral sessions in a two-dimensional probability distribution plot (Figure S1), (3) binned the distribution according to the value of $\mu(t_p)$, and (4) computed expected value of $\sigma(t_p)$ for each $\mu(t_p)$ bin. The larger circles with black outline correspond to grand averages of $\sigma(t_p)$ and $\mu(t_p)$ computed from combining t_p values for each trial type across all

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

behavioral sessions. The structure of variability (small circles connected with colored lines) was non-stationary, reward-dependent, and distinctly different from predictions of scalar variability (black line). Bottom: Relative reward rate is shown for each $\mu(t_p)$ bin of each trial type. The fraction rewarded was defined as the ratio between the number of rewarded trials and total number of trials across all behavioral sessions.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Deconstructing motor timing variability

The neurobiological basis of *scalar variability* is not understood. Models of interval timing have attributed scalar variability to a variety of stationary processes including a variable clock, a noisy accumulation process, noisy oscillations, and errors related to storage and/or retrieval of an interval (Church and Broadbent, 1990; Gibbon et al., 1984; Grossberg and Schmajuk, 1989; Jazayeri and Shadlen, 2010; Killeen and Fetterman, 1988; Machado, 1997; Oprisan and Buhusi, 2014; Simen et al., 2011; Staddon and Higa, 1999). According to all these models, $\sigma(t_p)$ should have a fixed linear relationship to $\mu(t_p)$. To test whether the noise processes were indeed stationary, we analyzed local estimates of $\mu(t_p)$ and $\sigma(t_p)$ across trials of the same type from blocks of 50 consecutive trials. The relationship between $\mu(t_p)$ and $\sigma(t_p)$ across blocks of trials was strikingly different from the predictions of scalar variability (Figure 1D, S1): local estimates of $\sigma(t_p)$ decreased when $\mu(t_p)$ was close to the desired t_t, and increased when $\mu(t_p)$ deviated from t_t. In other words, variability was smaller when the animal received reward and vice versa (Figure 1D). This inverse relationship was readily evident from the strong negative correlation between $\sigma(t_p)$ and the fraction of reward ($p << 0.001$, r = -0.48 for Monkey A and r = -0.50 for Monkey D).

Although this result rejects all models of scalar variability that assume stationarity, the fact that animals' behavior is modulated by reward is unsurprising. In our experiment, the only information animals could rely on to calibrate their behavior was reward. However, it is surprising that reward exerted its influence through adjustment of timing variability. This observation is consistent with the well-known explore-exploit strategy animals adopt when facing varying reward rates; i.e., explore (increase variability) when reward rate is low and exploit (reduce variability) when reward is certain (Dhawale et al., 2017; Pekny et al., 2015; Wu et al., 2014). Therefore, our results suggest that animals rely on reinforcement to actively calibrate their timing behavior. This highlights the need for more refined reinforcement learning models that can account for the dynamic effect of reinforcement on variability.

The non-stationarity of behavior was also evident from slow fluctuations of t_p throughout individual behavioral sessions (Figure 1B). These fluctuations have been reported in many tasks (Gilden et al., 1995; Merrill and Bennett, 1956; Weiss et al., 1955) and can be relatively strong in movements (Chaisanguanthum et al., 2014), reaction times (Laming, 1979), and interval timing (Chen et al., 1997; Murakami et al., 2017). To quantify these fluctuations, we analyzed the serial correlations of t_p as a function of the distance between the trials (i.e., trial lag) of the same type (e.g., all trials of ES within a session). We used partial correlation instead of simple correlation to avoid overestimating the temporal extent of the dependencies. In all four trial types, t_p exhibited significant serial correlations up to a trial lag of 20 or more ($p < 0.01$, dash lines: 1% and 99% confidence bounds by estimating the null distribution from shuffled series, Figure 2A).

One possible source of these correlations could be non-specific fluctuations of internal states such as arousal or fatigue (Murakami et al., 2017). For example, responses may exhibit slow

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

modulations if the animals' level of engagement or alertness were to wax and wane throughout the session. Since global internal state changes are nonspecific, a key prediction of this hypothesis is that slow fluctuations should persist across the four randomly interleaved trial types. We tested this prediction by comparing t_p correlations across the same and different trial types (different effectors and intervals). Serial correlations were stronger for trials with the same effector and interval (Figure 2B, S2) compared to trials of the same effector but different t_t (paired sample *t*-test on two sets of cross correlations with less than 20 trial lags and combining 4 trial types; monkey A: $p << 0.001$, $n = 80$, $t_{79} = 9.8$; monkey D: $p << 0.001$, $n = 80$, $t_{79} = 5.8$). In addition, correlations were stronger for trials of the same effector but different t_t compared to trials of different effectors (monkey A: : $p << 0.001$, $n = 80$, $t_{79} = 6.7$; monkey D: $p << 0.001$, $n = 80$, $t_{79} = 17.3$). The presence of stronger serial correlations across trials of the same type compared to trials of different type suggests that slow fluctuations of t_p were context-specific and cannot be fully explained in terms of global modulations of internal states.

Performance in CSG depends crucially on an accurate memory of t_t. Accordingly, we hypothesized that the slow fluctuations of t_p may reflect drifts in the animal's memory of t_t, which is thought to be a major contributor to motor timing error (Gibbon et al., 1984; Oprisan and Buhusi, 2014). To test this hypothesis, we reasoned that these fluctuations should be smaller if the demands on memory were reduced. Therefore, we trained a third monkey that had not been exposed to the CSG task to perform a variant of the task in which the interval t_t was measured on every trial, thereby minimizing the need to rely on memory (see Methods). As predicted, the behavior of the animal in this task did not exhibit any appreciable serial correlation (Figure S3) suggesting that the slow fluctuations in the original CSG task were a signature of drift in memory. For the remainder of our analyses, we will refer to these fluctuations as memory drifts.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

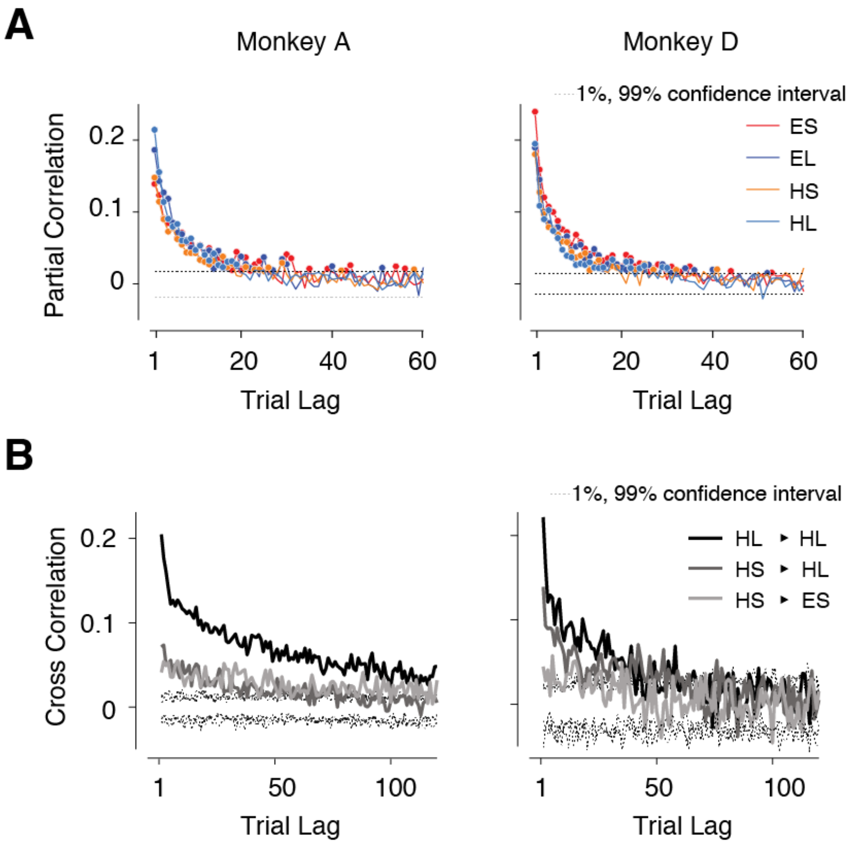

Figure 2. Context-dependent slow fluctuations of timing variability. (A) Long-term correlation of t_p across trials of the same type (same effector and target interval). For each behavioral session, and each trial type, we computed the partial correlation of t_p as the function of trial lag. Each curve shows the average partial correlations across all behavioral sessions. Eye-Short (ES), Eye-Long (EL), Hand-Short (HS), and Hand-Long (HL) trials are shown in different colors. Filled circles: correlation values that are larger than the 1% and 99% confidence interval (dashed line). (B) Comparison of long-term correlations of t_p across three example transitions between different trial types. The plot shows Pearson correlation coefficients as a function of trial lag. HL-HL: t_p correlation across trials transitioning between HL and HL trials; HL-HS: t_p correlation averaged across trials transitioning from HL to HS and from HS to HL. HS-ES: same for HS and ES trials. 1% and 99% confidence intervals were estimated from the null distribution. See Figure S2 for transitions between other conditions.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Reward regulates variability on a trial-by-trial basis

The drift of memory in CSG, if left unchecked, would hinder the animal's ability to perform the task. For example, a noisy random walk can generate slowly fluctuating outputs that undergo large excursions away from t_t (not shown). To maintain a reasonably accurate estimate of t_t and perform CSG, another process must counter the drift in memory. In CSG, the only information provided to calibrate memory is the reward outcome. Rewarded trials may be used to reinforce the memory while absence of reward may be construed as evidence that the memory is inaccurate and motivate animal to search for a different interval. Therefore, we hypothesized that the dependence of variability on reward is a signature of animals employing an explore-exploit strategy to continuously adjust their memory of t_t.

Our test of the nonstationarity of variability across blocks of trials (Figure 1D) provided initial support for this hypothesis: $\sigma(t_p)$ was small when $\mu(t_p)$ was close to desired t_t and reward probability was high. To further validate this hypothesis, we asked whether animals regulated their timing variability as a function of reward on a trial-by-trial basis. To do so, we analyzed the statistics of relative errors ($e = (t_p\text{- } t_t)/t_t$) across consecutive trials of the same type. We reasoned that error statistics should have two properties. First, due to drifts in memory, errors across consecutive trials should covary. Therefore, the average error in trial n, denoted $\mu(e^n)$, should be positively correlated with error in the preceding trial, denoted e^{n-1} (Figure 3A, top). Second, the explore-exploit strategy predicts that the standard deviation of errors, denoted $\sigma(e^n)$, should increase with the magnitude of e^{n-1} (Figure 3A, middle).

Deriving reliable estimates of $\mu(e^n)$ and $\sigma(e^n)$ as a function of e^{n-1} from a non-stationary process requires a large number of trials. Since our experiment consisted of four randomly interleaved trial types, the number of consecutive trials of the same type (e.g., ES followed by ES) within each behavioral session was limited. To address this limitation and gain statistical power, we combined our estimates of $\mu(e^n)$ and $\sigma(e^n)$ as a function of e^{n-1} across all trial types. We estimated $\mu(e^n)$ and $\sigma(e^n)$ using the following procedure: 1) we extracted all (e^{n-1},e^n) pairs for consecutive trials that were of the same type; 2) we collected all pairs in all the four trial types and grouped them into bins depending on the value of e^{n-1}; 3) we measured the mean and standard deviation of e^n for each bin of e^{n-1}.

Our first prediction was that $\mu(e^n)$ should increase monotonically with e^{n-1} for trials of the same type. The plot of $\mu(e^n)$ as a function of e^{n-1} was clearly consistent with this prediction (Figure 3B, top Figure S5A, B). To quantitatively test this prediction, we used linear regression ($\mu(e^n) = m_0+m_1e^{n-1}$), and found that the slope of the regression (m_1) was significantly positive when the consecutive trials were of the same type (Table 1).

Our second prediction was that $\sigma(e^n)$ should increase with the magnitude of e^{n-1} for trials of the same type. In other words, we expected the relationship between $\sigma(e^n)$ and e^{n-1} to be U-shaped. Again, results were consistent with this prediction (Figure 3B, middle Figure S5,B). To quantify this observation, we used quadratic regression ($\sigma(e^n) = s_0+s_1e^{n-1}+s_2(e^{n-1})^2$). For a U-shaped function the coefficient of the square term (s_2) must be positive. Moreover, if the U-shaped is

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

correctly centered near $e^{n-1} = 0$, the coefficient of the linear term (s_1) should be nearly zero. As predicted, we found that s_2 was significantly positive for trials of the same type and s_1 was close to zero (Table 1). Note that the choice of quadratic function was not theoretically motivated; we simply considered this to be an approximate function for testing the predicted convexity. Validation of the predictions about the relationship of $\mu(e^n)$ and $\sigma(e^n)$ with e^{n-1}, combined with the long-term serial correlations of t_p (Figure 2A) suggest that statistics of t_p can be understood in terms of two factors: 1) slow modulations of the mean of t_p reflecting drift in memory, and 2) modulations of the variability of t_p based on reward in the preceding trial.

Reward-dependent regulation of variability is context-specific

Our earlier analyses indicated that the slow memory drifts in the behavior were context-specific; i.e., dependent on trial type (Figure 2B). Therefore, for the explore-exploit strategy to be able to moderate drifts in memory, the effect of reward on variability must also be context-specific. That is, an increase in variability following a large error should only be present if the subsequent trial is of the same type. Otherwise, reinforcement of one trial type, say ES, may incorrectly adjust variability in the following trials of another type, say HL, and would interfere with the logic of harnessing reward to calibrate the memory of t_t.

To test this possibility, we performed the same analysis of $\mu(e^n)$ and $\sigma(e^n)$ as a function of e^{n-1} for pairs of consecutive trials associated with different effectors. Remarkably, the systematic relationship of both $\mu(e^n)$ and $\sigma(e^n)$ to e^{n-1} was nearly abolished (Figure 3B, open circles). Moreover, both the linear term of a linear regression model relating $\mu(e^n)$ to e^{n-1} and the quadratic term of a quadratic regression model relating $\sigma(e^n)$ to e^{n-1} were significantly reduced compared to data associated with same trial types (Table 1). As a final verification, we applied the same analysis across all pairs of trial types (ES vs ES, ES vs EL, etc.; Figure S5A,B). Results indicated that (1) correlations were strongest between pairs of trials of the same type (as expected from our previous partial correlation analysis in Figure 2A), and (2) the modulation of variability was most strongly and consistently present across trials of the same type. Since our analysis was based on data combined across sessions, one potential concern we had was that modulation of t_p variance with reward was due to differences across (but not within) sessions. To address this, we applied the same analysis to normalized errors derived from t_p values that were z-scored within each trial type and each session. This normalization scheme ensured that error pairs (e^n, e^{n-1}) in every session were drawn from a zero-mean and unit-variance distribution. Results of this complementary analysis validated our original inference that larger errors were associated with larger variances in the succeeding trial of the same but not other type (Figure S5C). Together, these results provide strong evidence that animals used reward in each trial to regulate behavioral variability in the next trial in accordance with an explore-exploit strategy and in a context-dependent manner.

The influence of reward on variability can be interpreted in two ways. One interpretation is that reward-dependent learning helps the system reduce variability (Krakauer and Mazzoni, 2011). An alternative interpretation is that, in the absence of reward, the system actively increase variability to explore the response space and find a rewarding solution (Dhawale et al., 2017).

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Generally, it has been difficult to furnish evidence that supports one interpretation over the other because there is no baseline with respect to which one can compare reward-dependent variability. Our experiment, which includes consecutive trials of both the same type and of different types may be able to shed some light on this question. We found that both the long-term serial correlations of t_p and the reward-dependent modulation of variability were greatly reduced when consecutive trials were of different types (Figure 2B, 3B). We assumed that this cross-condition provides a reasonable estimate of the baseline variability that is not sensitive to reward. We then compared the variability between consecutive trials of the same type both with and without reward to the cross-condition baseline. For both monkeys, $\sigma(e^n)$ was similar across the two conditions when e^{n-1} was relatively small but significantly larger for trials of the same type when e^{n-1} was large (Figure 3B; Table 1). This is consistent with the interpretation that animals acted more variably after unrewarded trials.

Reward-dependent context-specific regulation of variability in humans

To ensure that our conclusions were not limited to data collected from highly trained monkeys, we performed a similar experiment in human subjects. In human psychophysical experiment, t_t varied from session to session, and subjects had to constantly adjust their t_p by trial-and-error. Like monkeys, human behavior exhibited long-term serial correlations (Figure S4A), and these correlations were context (effector) specific (Figure S4B). We performed the same analysis as in monkeys to characterize the dependence of $\mu(e^n)$ and $\sigma(e^n)$ on e^{n-1}. Results were qualitatively unchanged: $\mu(e^n)$ increased monotonically with e^{n-1} verifying the presence of slow drifts, and $\sigma(e^n)$ had a U-shaped with respect to e^{n-1} indicating that subjects used the feedback to regulate their variability (Figure 3C and Table 1). Finally, similar to monkeys, the effect of reward on variability was context-specific (Figure 3C). This result suggests that the memory of a time interval is subject to slow drifts, and that humans and monkeys use reward-dependent regulation of variability as a general strategy to counter these drifts and improve performance.

One difference between human and monkey was the way in which reward or lack thereof altered variability. In monkeys, comparison between trials of the same effector and different effectors suggested that variability was increased in the absence of reward. In humans, on the other hand, variability was reduced after positively reinforced trials. The reason monkeys – not humans – employed a strategy based on increasing variability in the absence of reward might have been due to the fact that animals were trained for months or even years, and thus has a much smaller baseline variability compared to that in humans (compare Figure 3B middle to 3C middle, Table 1). With such low baseline variability, animals could effectively accommodate a strategy based on increasing variability.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

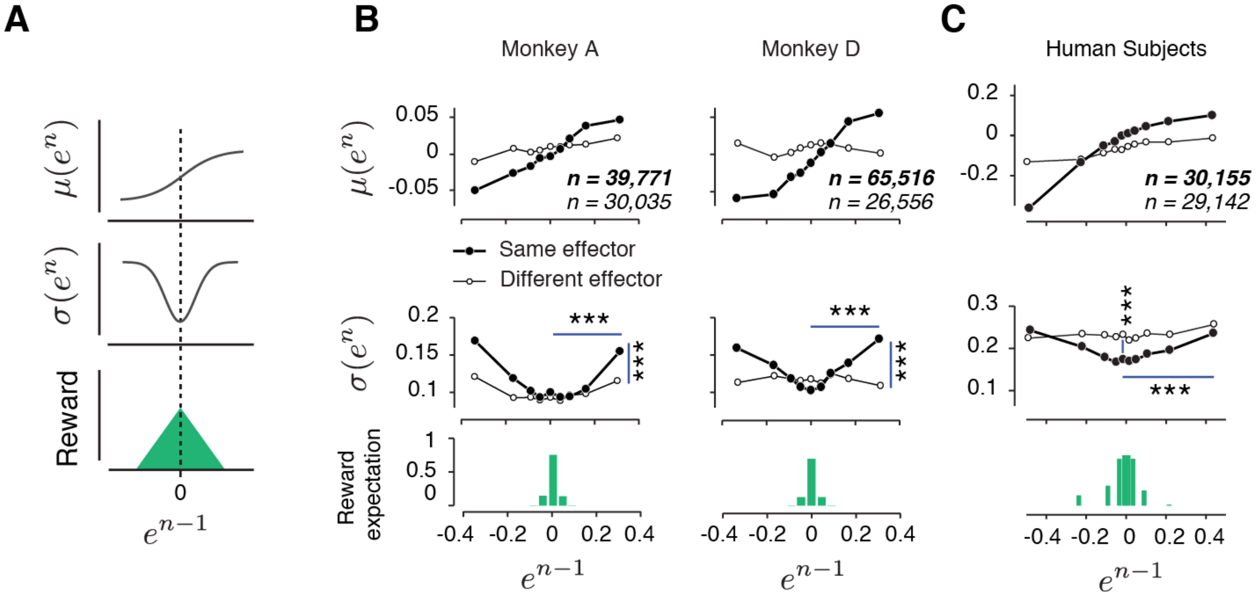

Figure 3. Rapid and context-dependent modulation of behavioral variability by reward. (A) Illustration of the expected effect of serial correlations and reward-dependent variability. Top: Positive serial correlations between produced intervals (t_p) creates a positive correlation between consecutive errors, and predicts a monotonic relationship between the mean of error in trial *n*, $\mu(e^n)$, and the value of error in trial *n-1* (e^{n-1}). Middle: Variability decreases with reward and increases in the absence of reward. This predicts a U-shaped relationship between the standard deviation of e^n, $\sigma(e^n)$, and e^{n-1}. Bottom: Reward as a function of e^{n-1}. (B) Trial-by-trial changes in the statistics of relative error. Top: $\mu(e^n)$ as a function of e^{n-1} in the same format shown in (A) top panel. Filled and open circles correspond to consecutive trials of the same and different types, respectively. Middle: $\sigma(e^n)$ as a function of e^{n-1}, sorted similarly to the top panel (Figure S4B shows similar results for each condition separately). Bottom: the reward expectation as a function of e^{n-1}. The reward expectation was computed by averaging the magnitude of reward received across trials. For humans, the reward was defined as the ratio between number of trials with positive feedback and total number of trials. In the same effector, variability increased significantly after an unrewarded trials compared to a rewarded trials (horizontal line, two-sample F-test for equal variances, *** p << 0.001) for both large positive errors (Monkey A: F(11169,10512) = 1.09, Monkey D: F(18540,13478) = 1.76) and large negative errors (Monkey A: F(8771,9944) = 1.40, Monkey D: F(21773,14889) = 1.62). The variability after an unrewarded trial of the same effector was significantly larger than after an unrewarded trial of the other effector (vertical line, two-sample F-test for equal variances, *** p << 0.001) for both large positive errors (Monkey A: F(11169,8670) = 1.20, Monkey D: F(18540,7969) = 1.32) and large negative errors (Monkey A: F(8771,5994) = 1.26, Monkey D: F(21773,7179) = 1.27). (C) Same as (B) for human subjects. In humans, the variability was also significantly larger after a negatively reinforced trial compared to a positively reinforced trial (horizontal line, two-sample F-test for equal variances, *** p << 0.001) for both large positive errors (F(5536,5805) = 1.19) and large negative errors (F(9366,9444) = 1.11). The variability after a positively reinforced trial of the same effector was significantly lower than after a positively reinforced trial of the other effector (vertical line, two-sample F-test for equal variances, *** p << 0.001, F(14497,15250) = 1.10).

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

	Monkey A	Monkey D	Humans
	Same vs. different effector	Same vs. different effector	Same vs. different effector
m_1	**0.16 [0.12, 0.19] > 0.04 [0.03, 0.06]**	**0.20 [0.12, 0.28] > -0.00 [-0.05, 0.04]**	**0.50 [0.27, 0.44] > 0.15 [0.09, 0.20]**
m_0	0.00 [-0.00 ,0.01] ~ 0.01 [0.01, 0.01]	-0.00 [-0.02, 0.01] ~ 0.01 [0.00, 0.02]	-0.03 [-0.08, 0.02] ~ -0.06 [-0.08 ,-0.05]
s_2	**0.50 [0.39, 0.60] > 0.24 [0.14, 0.33]**	**0.48 [0.19, 0.78] > -0.06 [-0.17 , 0.05]**	**0.31 [0.21, 0.42] > 0.06 [-0.03, 0.16]**
s_1	-0.02 [-0.04, 0.01] ~ 0.02 [-0.00, 0.04]	0.03 [-0.03 ,0.10] ~ -0.01 [-0.04, 0.01]	0.00 [-0.03, 0.04] ~ 0.03 [-0.00 , 0.06]
s_0	0.09 [0.09, 0.10] ~ 0.08 [0.08, 0.92]	0.11 [0.10, 0.12] ~ 0.12 [0.11 , 0.12]	**0.17 [0.16, 0.19] < 0.22 [0.22, 0.24]**

Table 1. **Quantitative assessment of the dependence of $\mu(e^n)$ and $\sigma(e^n)$ on e^{n-1}.** We used linear regression ($\mu(e^n) = m_0+m_1e^{n-1}$) to relate $\mu(e^n)$ to e^{n-1}, and quadratic regression ($\sigma(e^n) = s_0+s_1e^{n-1}+s_2(e^{n-1})^2$) to relate $\sigma(e^n)$ to e^{n-1}. Fit parameters and the corresponding confidence intervals [1%, 99%] are tabulated for each monkey and for humans. We compared the magnitude of fit parameters between the same versus different effector conditions (bold: significantly different).

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A generative model linking multiple timescales of variability to reward-based learning

Our analysis of animals' behavior rejected previous models of stationary and scalar noise as the source of timing variability, and revealed instead that this variability can be characterized in terms of two key factors: long-term serial correlations due to memory drift and fast trial-by-trial modulations due to reward. We aimed to develop a model that could emulate these effects and capture the behavior. Initially, we considered two classes of models: autoregressive models that readily capture serial correlations (Wagenmakers et al., 2004) and reinforcement learning models that explain how reward can guide behavior through a process of trial and error (Kaelbling et al., 1996; Sutton and Barto, 1998). However, these two are generally incompatible: classic autoregressive models are insensitive to reward, and reinforcement learning models cannot accommodate serial correlations of behavior that are unrelated to reward. Moreover, most current formulations of reinforcement learning are geared towards problems in which the state or action space is discrete such as multi-armed bandit problems (Dayan and Daw, 2008). When the domain of decisions is discrete, an unrewarded outcome would promote exploring other options. This problem has been studied extensively in reinforcement learning literature and is typically formulated in terms of explore-exploit algorithms such as ε-greedy and softmax (Sutton and Barto, 1998). However, generalization of explore-exploit strategies to cases when the variable of interest is continuous, as in our motor timing task, is not straightforward. Exploiting the current option would mean producing the same t_p, which is impossible, and exploring a new option is not well defined as there are infinite options to choose from. However, explore-exploit strategies can be readily generalized to continuous variables if we formulate the problem in terms of variability (Dhawale et al., 2017). An exploit strategy would demand a reduction in variability, while an explore strategy would manifest itself as increased variability. Indeed, this formulation matches the patterns of behavioral variability with respect to reward in CSG (Figure 1D, 3B, and 3C).

To simultaneously capture both the serial correlations of t_p and the dependence of t_p variability on reward, we developed a generative Gaussian process (GP) model. This choice was motivated by three factors: 1) GPs automatically describe observations up to their second order statistics, which are the relevant statistics in our data, 2) GPs offer a nonparametric Bayesian fit to long-term serial correlations, and 3) as we will describe, GPs can be readily augmented to implement reinforcement learning in a continuous state space.

GPs are characterized by a covariance matrix – also known as the GP kernel – that specifies the degree of dependence between samples, and thus determines how slowly the samples fluctuate. The most common and mathematically convenient formulation of this kernel is known as the "squared exponential" kernel function, denoted $\boldsymbol{K}_{SE}$ (Figure 4A, top left):

$$K_{SE}(n, n-r) = \exp(-\frac{r^2}{2l_{SE}^2})$$

In $\boldsymbol{K}_{SE}$, the covariance between any two samples (indexed by n and n-r) drops exponentially as a function of temporal distance between them (r) and the rate of drop is specified by the

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

characteristic length parameter, l_{SE}. When l_{SE} is small, samples are relatively independent, and when it is large, samples exhibit long range serial correlations (Figure 4A, left).

This GP, however, generates samples drifting away from t_t in a manner that is qualitatively different from behavioral data (Figure 4A, 2nd row, left). To capture the effect of reward, we devised an augmented model, which we refer to as the reward-sensitive GP model (RSGP). The full kernel for the RSGP model ($\boldsymbol{K}_{RSGP}$) is a weighted sum of two kernels, a classic squared exponential kernel ($\boldsymbol{K}_{SE}$) scaled by σ^2_{SE}, and a reward-sensitive kernel ($\boldsymbol{K}_{RS}$) scaled by σ^2_{RS} (Figure 4A, top). A third diagonal matrix ($\sigma^2_0\boldsymbol{I}$) was also added to adjust for baseline variance:

$$K_{RSGP}(n,n-r) = \sigma^2_{SE}K_{SE}(n,n-r) + \sigma^2_{RS}K_{RS}(n,n-r) + \sigma^2_0 I$$

$\boldsymbol{K}_{RS}$ also has the form of squared exponential with a length parameter, l_{RS}. However, to make $\boldsymbol{K}_{RS}$ reward-sensitive, the covariance terms were non-zero only for rewarded trials (Figure 4A, 1st row, middle). The reward was a binary variable determined by an acceptance window around t_t. This allowed past rewarded samples to have a higher leverage on future samples (i.e., higher covariance), and this effect dropped exponentially for rewarded trials farther in the past.

$$K_{RS}(n,n-r) = \begin{cases} \mathbf{exp}(-\frac{\boldsymbol{r}^2}{\mathbf{2}\boldsymbol{l}^2_{RS}}) & \text{if trial } \boldsymbol{n-r} \text{ was rewarded} \\ \mathbf{0} & \text{otherwise} \end{cases}$$

Intuitively, RSGP operates as follows: $\boldsymbol{K}_{SE}$ with a relatively large length parameter captures the long-term covariance across samples (Figure 4A, 3rd row, left). $\boldsymbol{K}_{RS}$ with a smaller length parameter regulates shorter-term covariances (Figure 4A, 3rd row, middle) and allows samples to covary more strongly with recent rewarded trials (Figure 4A, bottom, middle). Using simulations of the full model with $\boldsymbol{K}_{RSGP}$ as well as reduced models with only $\boldsymbol{K}_{SE}$ or $\boldsymbol{K}_{RS}$ (Figure 4A, 2nd row), we verified that both kernels were necessary and that the full RSGP model was capable of capturing both the slow fluctuations and the reward-dependent control of the variance (Figure 4A, 4th and 5th row). Moreover, we used simulations to verify that parameters of the model were identifiable; i.e., fits of the model to simulated data accurately recovered the ground truth parameters (Supplementary Table 1).

We fitted the RSGP model to behavior of both monkeys and humans (Figure 4B, S6). As predicted by our hypothesis, the fits to the characteristic length associated with serial correlations (l_{SE}) were invariably larger than that of the reward-dependent kernel (l_{RS}) (Monkeys: l_{SE} = 20.6 ± 21.4, l_{RS} = 2.0 ± 0.7; Humans: l_{SE} = 19.2 ± 21.8, l_{RS} = 1.3 ± 0.4; Median ± MAD). The model fit of variances (σ_{SE}, σ_{RS} and σ_0) are shown in Figure S6C. In monkeys, σ_0 and σ_{SE} but not σ_{RS} were significantly different between two effectors (σ_0: $p << 0.001$, one-tail two sample t-test, df = 482, t = 5.26; σ_{SE}: $p << 0.001$, two sample t-test, df = 482, t = 5.06; σ_{RS}: p = 0.13, t-test, df = 261, t = 1.5 for the Short interval, and p = 0.26, t-test, dt = 219, t = 1.1 for the Long interval). The dependence of σ_0 and σ_{SE} on effector was consistent with our session-wide analysis of variance (Figure 1C). In the human subjects, variance terms were more similar between effectors (p = 0.03 for σ_0, p = 0.01 for σ_{SE}, and p = 0.027 for σ_{RS}, two sample t-test, df = 118). Importantly,

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

across both monkeys and humans, the model was able to accurately capture the relationship of $\mu(e^n)$ and $\sigma(e^n)$ to e^{n-1} (Figure 4C). These results validate the RSGP as a candidate model for simultaneously capturing the slow fluctuations of t_p and the effect of reward on t_p variability.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

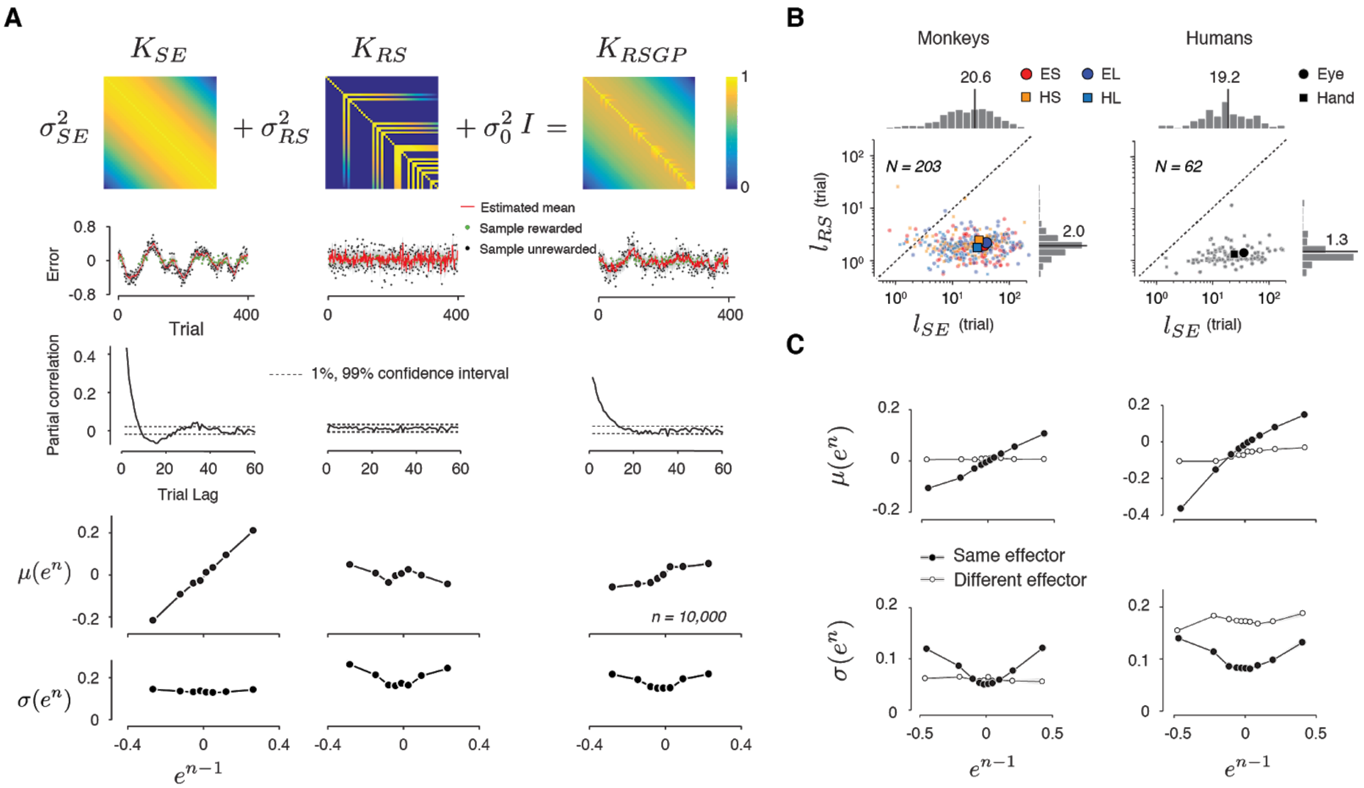

Figure 4. A reward-sensitive Gaussian process model (RSGP) capturing reward-dependent control of variability. (A) Top: The covariance function for the RSGP model (***K***$_{RSGP}$) is the sum of a squared exponential kernel (***K***$_{SE}$), a reward-dependent squared exponential kernel (***K***$_{RS}$) and an identity matrix (***I***) weighted by σ_{ES}^2, σ_{RS}^2, and σ_0^2, respectively. Second row: Simulations of three Gaussian process (GP) models, one using ***K***$_{ES}$ only (left), one using ***K***$_{RS}$ only (middle), and one with the full ***K***$_{RSGP}$ (right). Third row: Partial correlation of samples from the three GPs in the second row. Fourth and fifth row: The relationship between the mean (fourth row) and standard deviation (fifth row) of e^n as a function of e^{n-1} in previous trial, shown in the same format as in Figure 3B. Only the model with full covariance function captures the observed behavioral characteristics. (B) Length scales, l_{SE} and l_{RS} associated with ***K***$_{ES}$ and ***K***$_{RS}$, respectively, derived from fits of RSGP to behavioral data for monkeys (left) and humans (right). Small and large symbols correspond to individual sessions and the median across sessions, respectively. Different trial types are shown with different colors (same color convention as in Figure 1B). l_{RS} was significantly smaller than the l_{SE} (monkeys: $p << 0.001$, one-way ANOVA, $F_{1,945} = 463.4$; humans: $p << 0.001$, one-way ANOVA, $F_{1,235} = 102.5$). (C) Statistics of the predicted behavior from the RSGP model fits, shown in the same format as Figure 3B,C. Data were generated from forward prediction of the RSGP model fitted to behavior (see Methods for details). The standard error of the mean computed from n = 100 repeats of the average across bootstrap sample of trials is shown as shaded area but it is small and difficult to visually discern.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Slow fluctuations in thalamus

Recently, we identified a monosynaptic thalamocortical pathway from the medial portion of the ventral lateral thalamus, also known as area X, to the dorsomedial frontal cortex (DMFC) that plays a causal role in animals' flexible timing behavior (Wang et al., 2018). We also characterized the neural correlates of motor timing in this pathway: production of different time intervals was accompanied by temporal scaling of neural response profiles in DMFC, and the degree of scaling was predicted by the level of activity of antidromically identified DMFC-projecting neurons in thalamus. Using population data analysis and recurrent neural network modeling, we were able to explain these observations in terms of a simple two-step process: 1) animal's memory of the interval sets the firing rate of thalamic neurons ; 2) thalamic signals serve as a speed command and control the speed at which activity in DMFC evolve toward an action-triggering state. Based on these results, we reasoned that neural activity in thalamus may serve as a readout for drifts in animal's memory of t_t.

To test this, we recorded from multiple thalamic neurons around the region where direct thalamocortical projection was identified, and asked whether the activity across the population underwent slow fluctuations in register with t_p. Numerous studies have found that movement initiation time is predicted by signals that are established early during planning (Carpenter and Williams, 1995; Churchland et al., 2006; Hauser et al., 2018; Jazayeri and Shadlen, 2015; Lara et al., 2018; Remington et al., 2018a). Accordingly, we measured spike counts of simultaneously recorded thalamic neurons, denoted $\boldsymbol{r}_{\mathrm{Th}}$, within a 250 ms window before Set (Figure 5A, Left) and formulated a multi-dimensional linear regression model to examine the trial-by-trial relationship between $\boldsymbol{r}_{\mathrm{Th}}$ and slow fluctuations in error, which we denote by e_{slow} (Figure 5A, Right, red line). To solve the regression, we first needed to estimate e_{slow} on a trial-by-trial basis. To do so, we relied on the RSGP model, which readily decomposed the t_p time series to its slow memory-dependent and fast reward-dependent components. For each session, we fitted the RSGP to the entire session and then inferred the value of e_{slow} for each trial in that session (see Methods).

We computed the regression weight, $\boldsymbol{\beta}_{\mathrm{Th}}$, that when multiplied by $\boldsymbol{r}_{\mathrm{Th}}$, would provide the best linear fit to e_{slow}. If we think of the vector of spike counts in each trial as a point in a coordinate system where each axis corresponds to one neuron ("state space"), we can view $\boldsymbol{\beta}_{\mathrm{Th}}$ as a direction along which modulations of spike counts most strongly reflect memory drifts. Accordingly, we will refer to the direction associated with $\boldsymbol{\beta}_{\mathrm{Th}}$ as the drift direction, and will denote the strength of activity along that direction by z_{Th} ($z_{\mathrm{Th}} = \boldsymbol{r}_{\mathrm{Th}}\boldsymbol{\beta}_{\mathrm{Th}}$, Figure 5A, Right, blue line). To ensure that the model was predictive and not simply overfitting noisy spike counts, we used a cross-validation procedure: for each session, we used a random half of the trials to estimate $\boldsymbol{\beta}_{\mathrm{Th}}$, and the other half to quantify the extent to which z_{Th} could predict e_{slow}. Using this procedure, we found that in 91% of the sessions (201/220, all trial type combined), z_{Th} was significantly correlated with e_{slow} (Figure 5B, ** $p < 0.01$, null hypothesis test by shuffling trials)

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

suggesting that thalamic responses shortly before Set accurately reflected memory drifts in behavior.

Next, we analyzed the neurally-inferred drift (z_{Th}) across pairs of consecutive trials using the same approach we applied to behavioral data (Figure 3B, top). Specifically, we extracted pairs of ($e^{n\text{-}1}$, z^{n}_{Th}) for consecutive trials of the same type, binned them depending on the value of $e^{n\text{-}1}$, and measured $\mu(z^{n}_{\mathrm{Th}})$ for each bin of $e^{n\text{-}1}$. As expected, $\mu(z^{n}_{\mathrm{Th}})$ increased monotonically with $e^{n\text{-}1}$ as evidenced by the slope of a linear regression model relating $\mu(z^{n}_{\mathrm{Th}})$ to $e^{n\text{-}1}$ (filled circles in Figure 6B top, Table 2). Moreover, this relationship was absent for consecutive trials associated with different effectors (open circles in Figure 6B top, Table 2). These results indicate that thalamic responses along the identified drift direction prior to Set exhibit context-specific serial correlations analogous to behavior (Figure 3B top). These results indicate that the population activity in thalamus along a drift direction underwent slow fluctuations in register with drifts in animal's memory of t_t.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

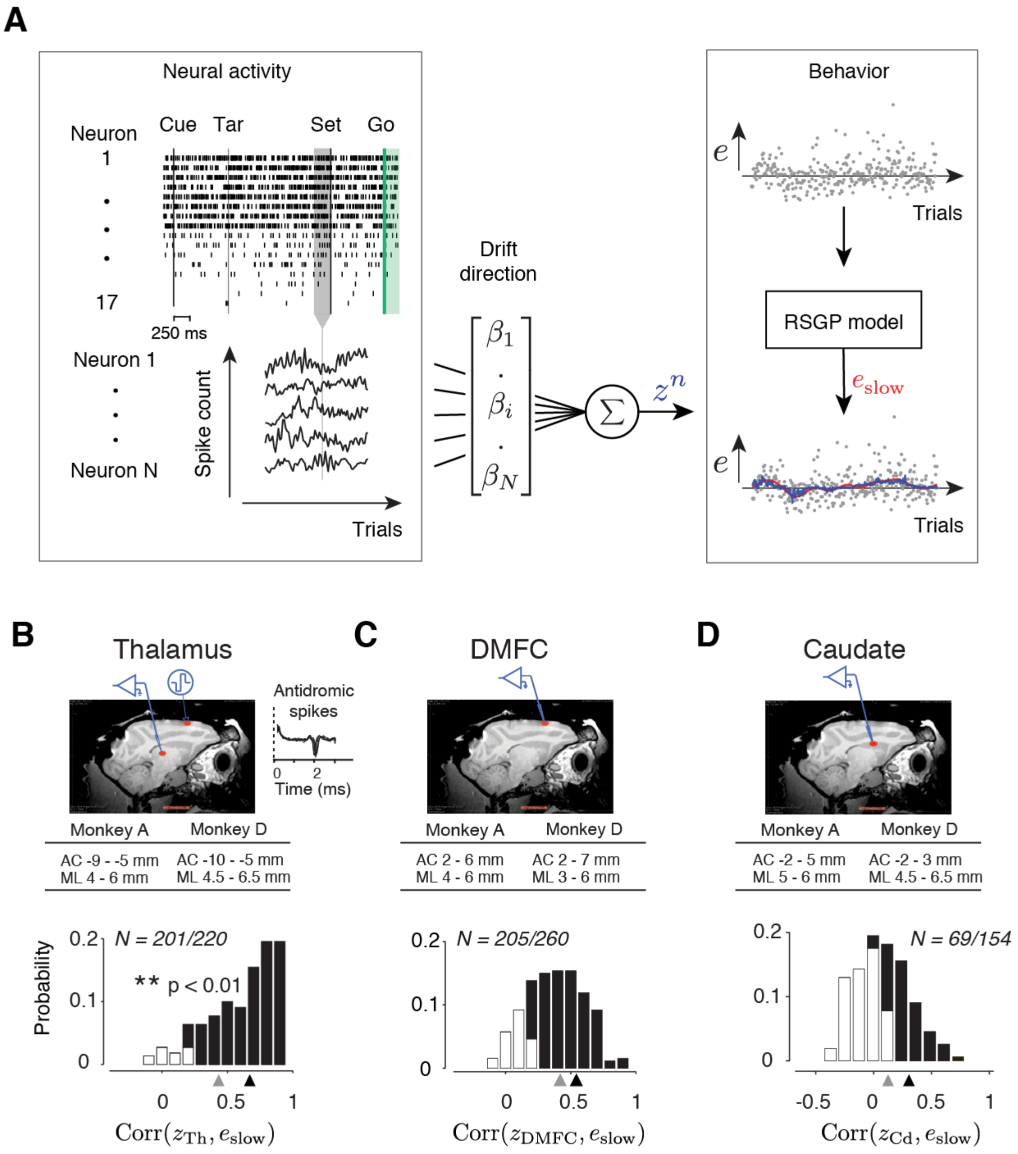

Figure 5. Representation of slow fluctuations of behavior in population activity. (A) The skeleton of the analysis for identifying the drift direction across a population of simultaneously recorded neurons. Top left: The rows show spike times (ticks) of 17 simultaneously recorded thalamic neurons in an example trial. From each trial, we measured spike counts within a 250 ms window before Set (gray region). Bottom left: The vector of spike counts from each trial (gray vertical line) was combined across trials providing a matrix containing the spike counts of all neurons across all trials. Middle: The matrix containing spike counts across trials and neurons was used as the regressor in a multi-dimensional linear regression model with weight vector, $\boldsymbol{\beta}$, to predict the slow component of error (e_{slow}). We refer to the direction of vector $\boldsymbol{\beta}$ as the drift direction across the population, and denote the projection of activity onto the drift direction on trial n by z^n. Right: We fitted the RSGP to errors (black dots, e), and then used the slow kernel of the fitted RSGP to infer e_{slow}. The plot shows the neurally inferred (z^n, blue) overlaid on e_{slow} derived from RSGP fit to behavior (red line). (B) Top: Parasagittal view of one of the animals (monkey D) with a red ellipse showing the regions targeted for electrophysiology. The corresponding stereotactic coordinates of the region of interest in each animal is tabulated (AC: anterior commissure; ML: mediolateral). Recorded thalamic neurons were from a region of thalamus with monosynaptic connections to DMFC (shown schematically by the stimulating electrode in DMFC). Inset:

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

antidromically activated spikes in thalamus. Bottom: Histogram of the correlation coefficients between $\boldsymbol{e}_{\text{slow}}$ inferred from the RSGP model and $\boldsymbol{z}^{n}_{\text{Th}}$ (projection of thalamic population activity on drift direction) across recording sessions. Note that some correlations are negative because of cross-validation (we used a random half of data to estimate the drift direction and the other half for estimation correlations). Otherwise, all correlations should have been non-negative. Black bars correspond to the sessions in which the correlation was significantly positive (**p < 0.01; hypothesis test by shuffling trial orders), and the remaining in white are those sessions in which correlation was not significantly different from that of bootstrapped data. The average correlation across all sessions and average correlation across sessions with significantly positive correlations are shown by gray and black triangles, respectively. (C) Same as B for DMFC. (D) Same as B for the caudate.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Reward-dependent regulation of variability in thalamus

Next, we sought to examine the neural underpinnings of our second key observation from behavior that variability increased after unrewarded trials (exploration) and decreased after rewarded trials (exploitation) in a context-specific manner. We hypothesized that reward regulates the variability of thalamic activity in the same manner that it regulates variability of t_p. Importantly, such reward-dependent regulation of neural variability must satisfy a crucial constraint: modulations to neural variability by reward should be restricted to the drift direction in the population activity (Figure 6A, bottom); otherwise this strategy will not be able to effectively counter the degrading effect of memory drift in a context-dependent manner.

In the previous section, we inferred the drift direction ($\boldsymbol{\beta}_{\mathrm{Th}}$) in thalamus from simultaneously recorded neurons. To test whether reward regulates variability along the drift direction, we examined the effect of reward on the variability of thalamic signals along the drift direction. Specifically, we analyzed $\sigma(z^n_{\mathrm{Th}})$ as a function of e^{n-1}, which is analogous to how we analyzed the effect of reward on behavioral variability (Figure 3B, middle). Remarkably, $\sigma(z^n_{\mathrm{Th}})$ as a function of e^{n-1} exhibited the characteristic U-shaped (Figure 6B bottom), which we verified quantitatively by comparing the variance of z^n_{Th} for rewarded and unrewarded trials ($p < 0.01$, two-sample F-test for equal variances on z^n_{Th} , F(5608,4266) = 1.08 for negative e^{n-1} and $p < 0.001$ for positive e^{n-1}, F(4723,6125) = 1.28). We further validated this result by verifying that the square term of a quadratic regression fit to the data ($\sigma(z^n_{\mathrm{Th}}) = s_0+s_1e^{n-1}+s_2(e^{n-1})^2$), was significantly positive (Table 2).

As an important control, we performed the same analysis between consecutive trials associated with different effectors and we found no significant relationship between $\sigma(z^n_{\mathrm{Th}})$ and e^{n-1} (Table 2, $p = 0.91$, two-sample F-test for equal variances, F(5332,3984) = 1.0 for the negative $\mathrm{e}^{\mathrm{n-1}}$; $p = 0.97$ for the positive $\mathrm{e}^{\mathrm{n-1}}$, F(4234,5857) = 0.99). Note that these analyses were repeated after square-root transforming spike count data to stabilize Poisson-like variability; this did not affect any of our conclusions. Taken together, these results indicate that variability of thalamic responses was modulated by reward, and that this modulation was aligned to the drift direction.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

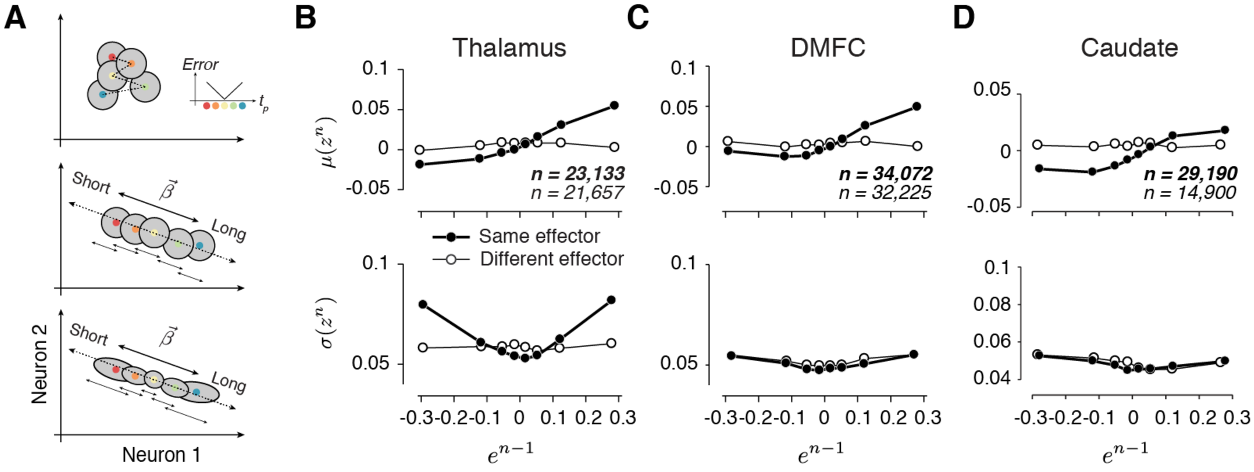

Figure 6. Alignment of reward-dependent neural variability and drift in thalamus, but not in DMFC and caudate. (A) Various hypotheses for how population neural activity could be related to produced intervals (t_p) shown schematically in 2 dimension (2 neurons). Top: The average neural activity (colored circles) is not systematically organized with respect to t_p, and the trial-by-trial variability of spike counts for a given t_p around the mean (gray area) is not modulated by the size of the error. The inset shows how error increases as t_p moves away from the target interval (t_t). Middle: Projection of average activity along a specific dimension (dotted line) is systematically ordered with respect to t_p, but the variability (small stacked arrows) does not depend on the size of error. Bottom: Projection of average activity along a specific dimension is systematically ordered with respect to t_p and the variability along the same axis increases with the size of error. (B) The relationship between neural activity in the thalamus on trial *n* to relative error in the preceding trial (e^{n-1}). Top: Expected mean of population activity on trial *n* ($\mu(z^n)$) along the drift direction ($\boldsymbol{\beta}$) as a function of e^{n-1}. Bottom: Expected standard deviation of population activity along the drift direction on trial *n* ($\sigma(z^n)$) as a function of e^{n-1}. Results are shown with the same format as in Figure 3B (filled circles: same effector; open circles: different effectors). (C) and (D) Same as (B) for population activity in DMFC and caudate. See Figure S7 for result of individual animal.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Parameters		Thalamus		DMFC		Caudate	
		Same effector	Different effector	Same effector	Different effector	Same effector	Different effector
$\mu(z^n)$	m_1	**0.14** [0.072, 0.20]	0.0075 [-0.022, 0.037]	**0.11** [0.0047,0.22]	-0.0039 [-0.028, 0.020]	**0.075** [0.023, 0.13]	0.0002 [-0.014, 0.015]
	m_0	0.0092[-0.001, 0.019]	0.006 [0.0014, 0.011]	0.0102 [-0.006, 0.027]	0.0071 [0.003, 0.011]	0.0046 [-0.0034, 0.013]	0.013 [0.011, 0.015]
$\sigma(z^n)$	s_2	**0.32** [0.24, 0.406]	0.0095 [-0.041, 0.060]	0.084 [-0.044, 0.130]	0.065 [-0.017, 0.11]	0.066 [-0.009, 0.12]	0.042 [-0.048, 0.13]
	s_1	0.0072 [-0.011, 0.025]	0.0021 [-0.009, 0.013]	0.0018 [-0.006, 0.011]	0.0016 [-0.008, 0.011]	-0.0063 [-0.018, 0.005]	-0.011 [-0.029, 0.004]
	s_0	0.055 [0.051, 0.058]	0.058 [0.056, 0.061]	0.047 [0.045, 0.048]	0.049 [0.047, 0.051]	0.046 [0.042, 0.047]	0.046 [0.042, 0.049]

Table 2. **Regression model fits relating spike count along the drift direction on trial *n* (z^n) to error in trial *n-1* (e^{n-1}).** m_0 and m_1 are parameters of the linear regression model relating the mean of z^n ($\mu(z^n)$) to e^{n-1}; i.e., $\mu(z^n) = m_0+m_1e^{n-1}$. s_0, s_1 and s_2 are parameters of the quadratic regression model relating the standard deviation of z^n ($\sigma(z^n)$) to e^{n-1}; i.e., $\sigma(z^n) = s_0+s_1e^{n-1}+s_2(e^{n-1})^2$. Fit parameters are shown separately for the thalamus, DMFC and caudate and further separated depending on whether trial *n-1* and *n* were of the same or different effectors. Bold values for m_1 and s_2 were significantly positive (** p < 0.01, 1% and 99% confidence intervals of the estimation were shown).

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Slow fluctuations and reward-dependent regulation of variability in DMFC and caudate

Our previous work indicated that thalamus provides the speed command at which DMFC responses evolve prior to movement initiation (Wang et al., 2018). Therefore, we extended our investigation of the neural signatures of slow fluctuations and reward-dependent regulation of variability to DMFC. To do so, we recorded from DMFC neurons simultaneously and examined responses using the same analysis we applied to the thalamus. We computed the regression weights, $\boldsymbol{\beta}_{\mathrm{DMFC}}$, that when applied to simultaneously recorded spike counts in DMFC, $\boldsymbol{r}_{\mathrm{DMFC}}$, would provide the best linear prediction of e_{slow}, denoted z_{DMFC} ($z_{\mathrm{DMFC}} = \boldsymbol{r}_{\mathrm{DMFC}}\boldsymbol{\beta}_{\mathrm{DMFC}}$). In ~79% of the sessions (205/260, all trial type combined), z_{DMFC} was significantly correlated with e_{slow} (Figure 5C; ** $p < 0.01$, null hypothesis test by shuffling trials) indicating DMFC responses also exhibited slow fluctuations in register with those observed in t_p. Moreover, z_{DMFC} exhibited monotonic increase in $\mu(z^n_{\mathrm{DMFC}})$ as a function of e^{n-1} (Figure 6C top, Table 2). Finally, this monotonic relationship was absent for consecutive trials associated with different effectors. Together, these results identified a context-dependent drift direction in DMFC ($\boldsymbol{\beta}_{\mathrm{DMFC}}$) along which responses exhibited serial correlations analogous to serial correlations of t_p. The presence of slow fluctuations in both the thalamus and DMFC is consistent with memory drift being a distributed process that impacts spiking activity across neural circuits involved in motor timing. Indeed, we have made similar observations in the caudate downstream of DMFC (Figure 5D, 6D and Table 2).

However, in DMFC and caudate, unlike thalamus, variability along the memory drift measured in terms of $\sigma(z^n_{\mathrm{DMFC}})$ and $\sigma(z^n_{\mathrm{Cd}})$, was relatively independent of e^{n-1} (Figure 6C bottom, 6D bottom). Indeed, in DMFC and caudate, we found no significant difference in standard deviation after rewarded and unrewarded trials (Table 2, two-sample F-test for equal variances, $F(6244,4818) = 0.87$, $p = 0.99$ for the negative e^{n-1}; $F(4825,7572) = 1.002$, $p = 0.021$ for the positive e^{n-1}, see Figure S6 for each animal separately). We further verified this observation by fitting $\sigma(z^n_{\mathrm{DMFC}})$ and $\sigma(z^n_{\mathrm{Cd}})$ as a function of e^{n-1} with a quadratic regression model, similar to what we did for $\sigma(z^n_{\mathrm{Th}})$. We found that the coefficient for the square term was not significantly different from zero (Table 2). These results indicate that, in DMFC and caudate, spiking activity aligned to $\boldsymbol{\beta}$, which provided a readout for the slow fluctuations of t_p, did not represent signals related to the reward-dependent regulation of behavioral variability. This lack of effect in DMFC and caudate serves as a negative control and has two important implications. First, it indicates that the alignment between the drift direction and the direction in which reward regulates variability is not a trivial consequence of our analysis, and second, it highlights the specificity of this alignment across the population of neurons in the thalamus.

Our analysis so far focused on spike counts in a fixed 250 ms time window before Set. Next, we applied the same analysis to different time points near the Cue, Set and Go events to investigate the regulation of neural variability throughout the trial. In thalamus, the effect of reward modulation could be traced back to the onset of Cue and persisted throughout the trial (Cue-Set: Figure 7 A left, Set-Go: Figure 7 A right; *** $p < 0.001$, two-sample F-test for equal variances, with the same method as in Fig 6 B bottom). In contrast, in DMFC and caudate,

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

reward had no consistent effect on neural variability throughout the trial (Figure 7 B,C). These results were consistent across the two monkeys (Figure S8). Together, these results suggest that reward exerts its influence on behavior by regulating the variability of the speed command provided by the thalamus.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

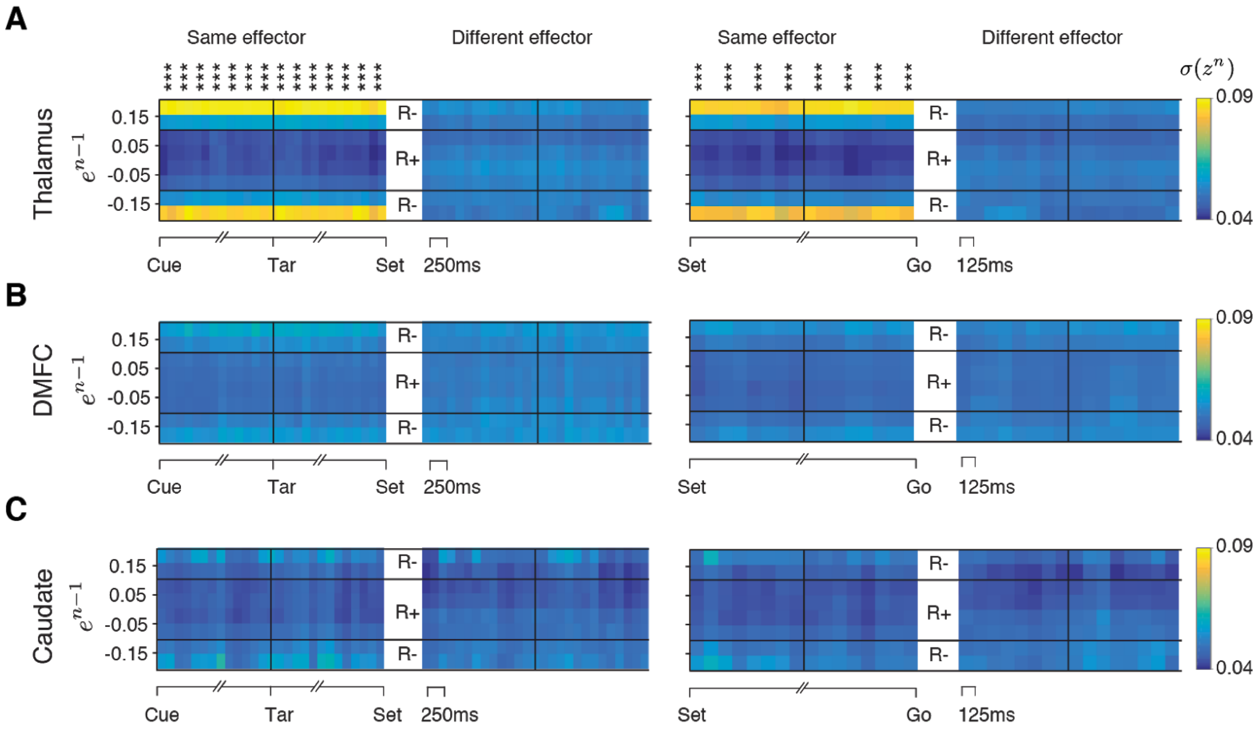

Figure 7. Analysis of reward-dependent neural variability throughout the trial. (A) Average standard deviation of population activity in thalamus at different points throughout the trial aligned to the Cue, Tar, Set and Go events. For each time point, we inferred the drift direction using the same analysis shown in Figure 5A, and projected spike counts onto the drift direction. We denote the projection of trial *n* by z^n. Each column shows the standard deviation of z^n ($\sigma(z^n)$) as a function of error in the preceding trial (e^{n-1}) based on spike counts within a 250 ms window centered at a particular time point in the trial. We grouped e^{n-1} in to 8 bins, 4 for negative e^{n-1} and 4 for positive e^{n-1}. The middle 4 bins are associated with reward (R+) in trial *n-1*, and the outer 4 bins, with no reward (R-) in trial *n-1*. Results are shown every 125 ms during the Cue-Set (left column) and Set-Go epochs (right column), separately for conditions in which tials *n-1* and *n* were of the same or different effectors (*** p< 0.001, using a two-sample F-test comparing the variance of z^n between rewarded and unrewarded trials separately for positive and negative values of e^{n-1}; H0: equal variance for at least one of the comparisons). (B) and (C) Same as (A) for DMFC and caudate. Results are for data combined across the two animals. Figure S8 shows the results of same analysis for each animal separately.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Discussion

Variability is a ubiquitous property of movements that can either degrade performance or be used for adaptive exploration to learn and improve performance. Here, we were able to advance our understanding of the function and neurobiology of variability along three axes. First, we found that the inherent behavioral variability in motor timing is comprised of a memory drift component that degrades performance and a reward-dependent exploratory component that improves performance. Second, we developed a novel reinforcement learning model that explains how reward or absence thereof can regulate timing variability. Finally, we characterized these sources of timing variability at the level of populations of neurons within the cortical-basal ganglia circuits involved in motor timing.

Several decades of research have characterized motor timing variability as a stationary process that scales with interval duration (Gallistel and Gibbon, 2000; Gibbon, 1977; Mauk and Buonomano, 2004). Our detailed analysis of behavior suggests that this variability is composite and contains at least two distinct non-stationary components. One component is characterized by slow fluctuations of behavior that result in serial correlations extending over minutes and across tens of trials, which corroborates previous reports in humans and animals (Chen et al., 1997; Gilden et al., 1995; Murakami et al., 2017). In most experiments, these fluctuations are attributed to fatigue, arousal or other nonspecific factors modulating internal state (Harris and Thiele, 2011; Kato et al., 2012; Lee and Dan, 2012; Luck et al., 1997; Niell and Stryker, 2010; Vinck et al., 2015). Although internal state modulations are likely present in our experiment, they do not seem to be the main driver of slow fluctuations. Nonspecific factors are expected to influence behavior in a nonspecific fashion. In contrast, we found that serial correlations in timing behavior were context-specific; i.e., they most strongly impacted trials associated with the same interval and same effector. Since different contexts in our experiment were randomly interleaved, the context-specificity of slow fluctuations cannot be straightforwardly explained in terms of global state changes.

Since knowledge of the desired time interval in our task is entirely dependent on memory, one hypothesis pertaining to serial correlations is that they reflect drifts in memory. We verified a critical prediction of this hypothesis in a control experiment in which memory demands were reduced, and found that indeed, the slow fluctuations were diminished. This result provides evidence in support of the conjecture that slow fluctuations in timing behavior are at least in part driven by drift in memory. However, the nature of biophysical and synaptic processes that lead to drifts in memory in a context-specific manner remains a fundamental and unresolved question. In our dataset, memory drifts exhibited relatively strong effector specificity. This suggests that the memory associated with producing an interval with different effectors is at least partially supported by distinct neural substrates. A more puzzling observation was the presence of weak but significant interval-specificity of memory drifts for the same effector. We don't have a definite explanation for this result but our previous work in the domain of time interval production (Wang et al., 2018) and reproduction (Remington et al., 2018a) as well as others studies in the motor system (Afshar et al., 2011; Ames et al., 2014; Hauser et al., 2018;

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Sheahan et al., 2016; Vyas et al., 2018) suggest that several aspects of movement control can be understood in terms of adjusting inputs and initial conditions of a dynamical system (Churchland et al., 2012; Remington et al., 2018b). Accordingly, the interval specificity of the memory drifts suggests that non-overlapping groups of neurons set the interval-dependent input and/or initial condition, which is consistent with our previous work (Wang et al., 2018). The degree to which behavioral contexts interfere may depend on the overlap between the corresponding neural representations. As such, the strong effector specificity and weaker interval specificity of the behavior predict a hierarchical organization of the underlying neural representations with maximum difference between effectors and smaller differences between intervals. Remarkably, this conjecture was corroborated by an analysis of the organization of Euclidean distances between activity patterns associated with the two effectors and intervals in all three brain areas (Figure S9). Together, these observations lead us to speculate that context-dependent memory drifts may be a general phenomenon in the motor system and that interaction between different contexts or movement variables may depend on the distance between the corresponding neural representations.

The other non-stationary component of motor timing variability became evident when we analyzed the behavior with respect to reward history. Variability was smaller after rewarded trials compared to unrewarded trials. This is a remarkable finding as it suggests that, for a behavioral task as simple as producing a time interval, the brain relies on a sophisticated reward-dependent control process that impacts variability. Previous work has noted the importance of regulating variability when there is need to learn and/or calibrate continuous behavioral parameters such as movement angle, speed or trajectory using an unsigned feedback such as presence or absence of reward (Pekny et al., 2015; Santos et al., 2015; Wu et al., 2014). This is directly relevant to out work since animals had to rely solely on reward to calibrate their memory of the desired interval. A crucial observation bolstering the link between variability and learning was that the effect of reward on variability was also context-specific; i.e., variability associated with producing a specific interval with a specific effector was most strongly dependent on reward history in trials of the same interval and effector. This observation serves as a strong indication that the function of the context-specific reward-dependent regulation of variability is to counter the corresponding memory drifts.

As previous work has noted (Dhawale et al., 2017; Fee and Goldberg, 2011; Pekny et al., 2015; Santos et al., 2015; Wu et al., 2014), modulation of variability by reward is broadly consistent with an explore-exploit strategy: search for other options when reward rate is decreasing and continue with the current option when reward rate remains high. This strategy plays a central role in models of reinforcement learning (RL). However, most existing RL models have focused on experimental settings in which the agent faces a discrete set of options such as multi-arm bandit tasks (Daw et al., 2006; Hayden et al., 2011; Lee et al., 2011; Wilson et al., 2014) for which the space of possible choices is limited. In these situations, RL models posit that the agent keeps track of the value of available options and adopts a suitable policy to choose among them (Sutton and Barto, 1998). However, these models cannot be straightforwardly adapted to continuous state spaces when the agent has to choose among infinite options. To

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

accommodate learning in the context of a continuous variable (i.e., time), we developed a non-stationary and reward-sensitive Gaussian process model (RSGP) capable of simultaneously accounting for the slow fluctuations of behavior across trials and reward-dependent adjustments of behavioral variability. RSGP can be viewed as a hybrid between autoregressive processes that are typically used to capture serial correlations in behavior and reinforcement learning models that account for how the agent uses feedback to update the value of options. As such, RSGP offers a general framework for future studies examining how feedback may be used to guard against ubiquitous nonstationarities in behavior (Chaisanguanthum et al., 2014; Chen et al., 1997; Gilden et al., 1995; Laming, 1979; Merrill and Bennett, 1956; Murakami et al., 2017; Weiss et al., 1955).

Intriguingly, the relative timescales of variability in CSG are comparable to those associated with a different type of motor learning based on sensory feedback, also referred to as error-based motor learning (Huberdeau et al., 2015; Smith et al., 2006; Wolpert et al., 2011). In both cases, the fast process reflects rapid learning either through reinforcement as we have found, or via sensory feedback in error-based learning. However, our interpretation of the slow process differs from error-based learning (Joiner and Smith 2008). We characterized the slow process as drifts in memory, whereas in error-based learning, this process is thought to play an active role in learning. However, it is conceivable that the slow component of timing variability also contributes to learning through an active averaging process that maintain a stable memory of the desired interval. Finally, we note that previous work in error-based motor learning focused on the two timescales of learning in the context of forming new memories. Our work extends the function of these two processes to trial-by-trial calibration required for maintenance of a stable memory.

Next, we examined the neurobiological underpinnings of memory drift and reward-dependent regulation of variability. The memory of a time interval is likely supported by distributed biophysical and synaptic mechanisms in the cortical and subcortical areas involved in motor timing (Gibbon et al., 1984; Mauk and Buonomano, 2004; Paton and Buonomano, 2018). Therefore, to gain a detailed mechanistic understanding of drifts of memory, one must measure and characterize various stochastic cellular mechanisms *in-vivo*, which is experimentally not feasible. With this limitation in mind, we reasoned that cellular processes that underlie such drifts may nonetheless impact the spiking activity of neurons that control animal's timing behavior.

Previously, we found a causal role for a cortico-basal ganglia circuit comprised of DMFC, DMFC-projecting thalamus and caudate in the control movement initiation time (Wang et al., 2018). We found that thalamic neurons provide an effector- and interval-dependent input that is integrated in DMFC to control the speed at which population activity in DMFC and the downstream caudate evolve toward a terminal movement-initiation state, i.e., faster for shorter intervals and slower for longer intervals (Wang et al., 2018). Therefore, we reasoned that

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

spiking activity within this circuit should exhibit features that are consistent with both context-dependent memory drifts and the reward-dependent adjustment of variability.

A regression analysis revealed that indeed the population activity in all three areas carry a signal correlated with drifts in memory. This finding is broadly consistent with previous studies reporting correlates of internal state changes and/or slow behavioral fluctuations in the thalamus (Halassa et al., 2014), the medial frontal cortex (Karlsson et al., 2012; Murakami et al., 2017; Narayanan and Laubach, 2008; Sul et al., 2010), and the caudate (Lau and Glimcher, 2007; Lauwereyns et al., 2002; Santos et al., 2015). However, the effector- and interval-dependent nature of these fluctuations in our data suggest that they may partially reflect context-specific memory drifts. The key feature of our task that enabled us to discover this specificity was the need to switch between different contexts on a trial by trial basis. Since many previous studies did not employ this feature, it is possible that, certain aspects of neural variability previously attributed to nonspecific internal state changes were in part caused by memory drifts related to task rules and parameters. Indeed, drifts and degradation of instrumental memories may be a key limiting factor in motor skill performance (Ajemian et al., 2013).

Although we found a correlate of these drifts in all three brain areas, we cannot make a definitive statement about the loci at which the underlying synaptic and biophysical drifts may be traced. It is likely that the memory has a distributed representation in which case the drift may also result from stochastic processes distributed across multiple brain areas. However, it is also possible that context information about effector and interval is stored in specific sub-circuits such as corticostriatal synapses (Fee and Goldberg, 2011; Xiong et al., 2015) and circuit-level interactions allow these drift to be manifest across other nodes of the cortico-basal ganglia circuit.

Next, we asked whether the variability of neural activity in the thalamus, DMFC and caudate were modulated by reward in the preceding trial in the same context-dependent manner as in the behavior. According to our hypothesis, the key function of the reward-dependent regulation of variability is to counter memory drifts. This hypothesis makes a specific and non-trivial predictions: reward should modulate the specific pattern of population neural activity that corresponds to memory drifts in the behavior, which we referred to as drift direction. Analysis of neural activity revealed that this effect was present in thalamus but not in DMFC and caudate. Indeed, only in thalamus, spike count variability increased after rewarded trials and decreased after unrewarded trials in a context-specific manner. Previous studies have reported that firing rates were modulated in thalamus on a trial-by-trial basis in the service of attention (Mcalonan et al., 2008; Saalmann et al., 2012; Zhou et al., 2016) and rule/context dependent computations (Schmitt et al., 2017; Wang et al., 2018). Our work demonstrates that modulations of thalamic activity may additionally subserve reward-based calibration of movement initiation times. It would be important for future studies to investigate whether this finding would generalize to other movement parameters.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The fact that the same effect was not present in DMFC and caudate strengthens our conclusions; it suggests that the reward-dependent regulation of variability in thalamus cannot be attributed to our analysis technique. However, this begs the question of why this regulation was not inherited by DMFC and caudate, especially by DMFC that receives direct projection from the region of thalamus we recorded from. The answer to this question depends on the nature of signal transformations along the thalamocortical pathway. While some experiments have suggested similar response properties for thalamus and their cortical recipients (Guo et al., 2017; Sommer and Wurtz, 2006), others have found that thalamic signals may undergo specific transformations along the thalamocortical pathway. For example, in early sensory areas, orientation selectivity is largely due to transformation at the LGN to V1 synapses (Hubel and Wiesel, 1962). Similarly, higher order thalamocortical inputs are thought to have a modulatory effect on cortical representations and dynamics (Berman and Wurtz, 2011; Schmitt et al., 2017; Wang et al., 2018; Wimmer et al., 2015). Therefore, the extent to which we should expect responses in DMFC and thalamus to have similar properties is unclear.

We previously found that the nature of signals in DMFC-projecting thalamus and DMFC during motor timing was indeed different: DMFC neurons had highly heterogeneous response profiles that evolved at different speeds depending on the interval, whereas thalamic neurons carried signals whose strength (i.e., average firing rate) encoded the underlying speed. We think that this transformation may provide an explanation for why reward-dependent modulation of firing rates was evident in thalamus but not in DMFC. Since thalamic neurons encode interval in their average firing rates, it is expected that regulation of timing variability by reward would similarly impact average firing rates. In contrast, in DMFC, the key signature predicting behavior was the speed at which neural trajectories evolved over time – not the average firing rates. This predicts that reward should alter the variability of the speed of neural trajectories. In principle, it is possible to verify this prediction by estimating the variance of the speed of neural trajectories as a function of reward. However, this estimation is challenging for two reasons. First, speed in a single trial is derived from changes in instantaneous neural states, and estimation of instantaneous neural states is unreliable unless the number of recorded neurons is far exceeds the dimensionality of the subspace containing the neural trajectory (Gao et al., 2017). Second, our predictions are about the variance – not mean – of speed, and estimating variance adds another layer of statistical unreliability unless the number of neurons or trials are sufficiently large.

Nonetheless, we devised a simple analysis to estimate the variance of speed of neural trajectories across single trials in all three areas. Results were consistent with our predictions: variance of speed was larger after unrewarded trials in DMFC and caudate but not thalamus, and this effect was present only for consecutive trials associated with the same effector (Figure S10). In other words, our results agree with the interpretation that reward controls the variance of average firing rates in thalamus, and this effect leads to the control of the variance of the speed at which neural trajectories evolve in DMFC and caudate.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

One question raised by our work is what brain areas supply the relevant information for the reward-dependent control of behavioral variability. While we cannot address this question definitively, we note that the area of thalamus we have recorded from receives information from three major sources, the frontal cortex , the output nuclei of the basal ganglia , and the deep nuclei of the cerebellum (Kunimatsu et al., 2018; Middleton and Strick, 2000). Therefore, one possibility is that the neural variability in the thalamus is adjusted by other cortical areas. Rapid reduction of variability prior to movement initiation has been found in motor and premotor areas (Churchland et al., 2006, 2010). Moreover, numerous experiments have reported trial-by-trial adjustment of correlated cortical variability in the domain of attentional control. In particular, it has been shown that spatial cueing leads to a drop of correlated variability across neurons whose receptive fields correspond to the cued location (Cohen and Maunsell, 2009; Mitchell et al., 2009; Ni et al., 2018; Ruff and Cohen, 2014). It is therefore possible that variability of firing rates in thalamus originates in cortex. However, to act as an effective mechanism in our motor timing task, correlated cortical variability must be additionally sensitive to reward-dependent neuromodulatory signals such as dopamine (Frank et al., 2009) possibly by acting on local inhibitory neurons (Huang et al., 2019). The basal ganglia could also play a role in reward-dependent control of thalamic firing rates (Kunimatsu and Tanaka, 2016; Kunimatsu et al., 2018). For example, single neuron responses in substantia nigra pars reticulata that are strongly modulated by reward schedule (Yasuda and Hikosaka, 2015) can influence neural responses in the thalamus. Finally, the cerebellum plays a central role in trial-by-trial calibration of motor variables (Herzfeld et al., 2015; Ito, 2002; Medina and Lisberger, 2008) including movement initiation time (Ashmore and Sommer, 2013; Kunimatsu et al., 2018; Narain et al., 2018) and thus is a natural candidate for calibrating firing rates in thalamus although how such calibration could be made reward-sensitive remains an open question (Hoshi et al., 2005). In sum, our work provides behavioral, modeling and neurophysiological evidence in support of the tantalizing hypothesis that the brain uses reinforcement to actively regulate behavioral variability in a task and context-dependent manner. This finding opens the possibility for future experiments to further investigate the underlying mechanisms.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Methods

Experimental model and subject details

Three adult monkeys (Macaca mulatta; one female, two males), and five human subjects (18-65 years, two females and three males) participated in this experiment. The Committee of Animal Care and the Committee on the Use of Humans as Experimental Subjects at Massachusetts Institute of Technology approved the animal and human experiments, respectively. All procedures conformed to the guidelines of the National Institutes of Health.

Animal experiments

Monkeys were seated comfortably in a dark and quiet room. 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 monitor (Acer H236HL, LCD) at a resolution of 1920x1080, and a refresh rate of 60Hz). Auditory stimuli were played from the computer's internal speaker. Eye position was 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.

The Cue-Set-Go task. Behavioral sessions in the main experiment consisted of four randomly interleaved trial types in which animals had to produce a target interval (t_t) of either 800 ms (Short) or 1500 ms (Long) using either a button press (Hand) or a saccade (Eye). The trial structure is described in the main Results (Figure 1A). Here, we only describe the additional details that were not described in the Results. The "Cue" presented at the beginning of each trial consisted of a circle and square. The circle had a radius of 0.2 deg and was presented at the center of the screen. The square had a side of 0.2 deg and was presented 0.5 deg below the circle. For the trial to proceed, the animal had to foveate the circle (i.e., eye fixation) and hold its hand gently on the button (i.e., hand fixation). The animal had to use the hand contralateral to the recorded hemifield. We used an electronic window of 2.5 deg around the circle to evaluate eye fixation, and infrared emitter and detector to evaluate hand fixation. After 500-1500 ms delay period (uniform hazard), a saccade target was flashed eight deg to the left or right of the circle. The saccade target ("Tar") had a radius of 0.25 deg and was presented for 250 ms. After another 500-1500 ms delay (uniform hazard), an annulus ("Set") was flashed around the circle. The Set annulus had an inner and outer radius of 0.7 and 0.75 deg and was flashed for 48 ms. Trials were aborted if the eye moved outside the fixation window or hand fixation was broken before Set.

For responses made after Set, the produced interval (t_p) was measured from the endpoint of Set to the moment the saccade was initiated (eye trial) or the button was triggered (hand trial). Reward was provided if the animal used the correct effector and t_p was within an experimentally controlled acceptance window. For saccade responses, reward was not provided if the saccade endpoint was more than 2.5 deg away from the extinguished saccade target, or if the saccade endpoint was not acquired within 33 ms of exiting the fixation window.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The width of the reward acceptance window was adjusted adaptively on a trial-by-trial basis and independently for each condition using a one-up one-down staircase procedure. As such, animals received reward on nearly half of trials (57% in monkey A and 51% in monkey D), and the magnitude of the reward scaled linearly with accuracy. Additionally, rewarded trials were accompanied by visual feedback and a brief auditory click. For the visual feedback, either the color of the saccade target (for Eye trials) or the central square (for the Hand trials) turned green.

No-memory control task. To validate our hypothesis that slow fluctuations in animals' behavior arose from memory fluctuations, we performed a control experiment in a third, naive monkey. In the control experiment**,** the animal did not have to remember the target interval t_t, but instead measured it on every trial. This was done by presenting an additional flash ("Ready") shortly before the Set flash such that the interval between Ready and Set was fixed and equal to t_t. This effectively removed the need for the animal to hold the target interval in memory. Based on our observations that slow fluctuations were context-dependent, we limited the control experiment to a single effector (Eye) and a single interval (t_t = 840 ms).

Electrophysiology. Recording sessions began with an approximately 10-minute warm up period to allow animals to recalibrate their timing and exhibit stable behavior. We recorded from 932 single- or multi-units in thalamus, 568 units in dorsomedial frontal cortex (DMFC), and 509 units in caudate, using 24-channel linear probes with 100 µm or 200 µm interelectrode spacing (V-probe, Plexon Inc.). The dorsomedial frontal cortex (DMFC) comprises supplementary eye field, dorsal supplementary motor area (i.e., excluding the medial bank), and pre-supplementary motor area. Recording locations were selected according to stereotaxic coordinates with reference to previous studies as well as each animal's structural MRI scan. The region of interest targeted in the thalamus was within 1 mm of antidromically identified neurons. All behavioral and electrophysiological data were timestamped at 30 kHz and streamed to a data acquisition system (OpenEphys). 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).

Antidromic Stimulation. We used antidromic stimulation to localize DMFC-projecting thalamic neurons. Antidromic spikes were recorded in response to a single biphasic pulse of duration 0.2 ms (current < 500 uA) delivered to DMFC via low impedance tungsten microelectrodes (100 – 500 KΩ, Microprobes). A stainless cannula guiding 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 3ms, with less than 0.2 ms jitter.

Human experiments

Each experimental sessions lasted approximately 60 minutes. Each subject completed 2-3 sessions per week. Similar to monkeys, experiments were conducted using the MWorks software. All stimuli were presented on a black background monitor. Subjects were instructed to

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

hold their gaze on a fixation point and hold a custom made push button throughout the trial. Subjects viewed the stimuli binocularly from a distance of approximately 67cm on a 23-inch monitor (Apple, A1082 LCD) driven by a Mac Pro at a refresh rate of 60 Hz in a dark and quiet room. Eye positions were tracked with an infrared camera (Eyelink 1000 plus, SR Research Ltd.) and sampled at 1 kHz. State of the button was converted and registered as digital TTL through a data acquisition card (National Instruments, USB-6212). The Cue-Set-Go task for humans was similar to monkeys with the following exceptions: (1) in each session, we used a single t_t sampled from a normal distribution (mean: 800ms, std: 80ms); (2) the saccadic target was 10 deg (instead of 8 deg) away from the fixation point. On average, 50.2% of trials received positive feedback.

Data Analysis

All offline data processing and analyses were performed in MATLAB (2016b, MathWorks Inc.).

Analysis of behavior

Behavioral data for the CSG task comprised of N = 203 behavioral sessions consisting of n = 167,115 trials in monkeys (N = 95, n = 71,053 for Monkey A and N = 108, n = 96,062 for Monkey D), and N = 62 sessions and n = 59,297 trials in humans (combined across subjects). Behavioral data for the no-memory control task was collected in N = 9 sessions and n = 32,041 trials in a third naive monkey.

We computed the mean and standard deviation of t_p, denoted by $\mu(t_p)$ and $\sigma(t_p)$, respectively, for each trial type within each session (Figure 1C). We additionally analyzed local fluctuations of $\mu(t_p)$ and $\sigma(t_p)$ by computing these statistics from running blocks of 50 trials within session and averaged across sessions. The resulting distribution of local $\mu(t_p)$ and $\sigma(t_p)$ were shown in Figure S1. The mean of $\sigma(t_p)$ for each corresponding $\mu(t_p)$ bin and the averaged reward across all trials in each $\mu(t_p)$ bin were plotted in Figure 1D. Results were qualitatively the same when the block length was increased or decreased by a factor of two.

We also examined the slow fluctuations of t_p for pairs of trials that were either of the same type (e.g., Eye-Short versus Eye-Short) or of different types (e.g., Hand-Long versus Eye-Short). For trials of the same type, we computed partial correlation coefficients of t_p pairs by fitting successive autoregressive model with maximum order of 60 trial lag (Box et al., 2015) (Figure 2A). 1% and 99% confidence bounds were estimated at 2.5 times the standard deviation of the null distribution. For trials of different types, we calculated the Pearson correlation coefficient of pairs of t_p of various lags (Figure 2B, Figure S2 and S4). To clarify our analysis, we use an example of how we estimated the cross correlation between pairs of HS - ES with a trial lag of 10: (1) normalize (z-score) two t_p vectors associated with HS and ES in each session; (2) take pairs of HS-ES that are 10 trials apart within each session; (3) combine the pairs across sessions; (4) compute Pearson correlation coefficient. We also computed a corresponding null distribution from 100 randomly shuffled trial identity. 1% and 99% confidence intervals were estimated from the null distribution.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Finally, we quantified the mean and standard deviation of the relative error denoted by $\mu(e^n)$ and $\sigma(e^n)$ as a function of error in the previous trial (e^{n-1}) for each pair of trial types (Figure S5 A and B). Since each trial can be of four different types (ES, EL, HS, HL), consecutive trials comprise 16 distinct conditions (e.g., ES-EH, HL-EL, etc, Figure S5A shows the distribution for all conditions). The limited number of trials in each session limited the reliability of statistics estimated for each individual condition. To gain more statistical power, we combined results across trials types in two ways. First, for each effector, we combined the Short and Long trials types by normalizing t_p values by their respective t_t. The resulting variable was defined as relative error e^n = (t_p^n- t_t)/t_t.This reduced the number of conditions by a factor of four, leaving consecutive trials that were either associated with the same effector or with different effectors (e.g., E-E, E-H, H-E, and H-H). We further combined trials to create a “same effector” condition that combined E-E with H-H, and a “different effector” condition that combined E-H with H-E. Animals and human subjects were allowed to take breaks during the experimental sessions. However, the pairs of consecutive trials used in all analyses, regardless of the trial condition, were restricted to the two consequent and completed trials that were no more than 7 second apart.

Reward-sensitive Gaussian process (RSGP) model simulation and fitting

We constructed a reward-sensitive Gaussian process model whose covariance function, $\boldsymbol{K}_{RSGP}$, is a weighted sum of two kernels, a traditional squared exponential kernel, for which we used subscript SE ($\boldsymbol{K}_{SE}$), and a reward-sensitive kernel with subscript RS ($\boldsymbol{K}_{RS}$). The two kernels contribute to K_{RSGP} through scale factors σ^2_{SE} and σ^2_{RS}, respectively. In both kernels, the covariance term between any two trials (trial n and n-r) drops exponentially as a function of trial lag (r). The rate of drop for $\boldsymbol{K}_{SE}$ and $\boldsymbol{K}_{RS}$ are specified by characteristic length parameters, l_{SE} and l_{RS}, respectively. The model also includes a static source of variance, $\sigma^2_0\boldsymbol{I}$ ($\boldsymbol{I}$ stands for the identity matrix)*:*

$$K_{RSGP}(n, n-r) = \sigma^2_{SE} K_{SE}(n, n-r) + \sigma^2_{RS} K_{RS}(n, n-r) + \sigma^2_0 I$$

$$K_{SE}(n, n-r) = \exp(-\frac{r^2}{2l^2_{SE}})$$

$$K_{RS}(n, n-r) = \begin{cases} \mathbf{exp}(-\frac{r^2}{2l^2_{RS}}) & \text{if trial } \boldsymbol{n-r} \text{ was rewarded} \\ \mathbf{0} & \text{otherwise} \end{cases}$$

Note that $\boldsymbol{K}_{RS}$ values depend on reward history and are thus not necessarily invariant with respect to time; i.e. $K(n, n-i) \neq K(m, m-i)$. This formulation allows past rewarded trials to have a higher leverage on future trials and this effect drops exponentially for rewarded trials farther in the past.

We simulated the RSGP by applying GP regression based on the designated covariance function (Table 3). To simplify our formulations and without loss of generality, we replaced σ^2_{SE} and σ^2_{RS} by $\alpha\sigma^2$ and $(1 - \alpha)\sigma^2$, respectively where α = 1.0, 0, and 0.5 for the three examples (Figure 3A, Table S1).

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

```
for $i := 1...n$ do
    1. Given the previous value and reward history, infer the mean and variance of $t_p^n$ from the conditional distribution
       $[tp^1, tp^{n-1}, tp^n]^\intercal \sim \mathcal{N}(t_t, \begin{bmatrix} K_{(n-1)\times(n-1)}, K_{(n-1)\times 1} \\ K_{1\times(n-1)}, K_{1\times 1} \end{bmatrix})$ , $K \equiv K_{RSGP}$
    2. Randomly sample $t_p^n$ from the inferred mean and variance
    3. Update the reward, reward-sensitive covariance $K_{RS}$, and hence $K_{\mathrm{RSGP}}$ based on $t_p^n$.
end for
```

Table 3. Algorithm for generating time series based on RSGP model.

As both the slow fluctuation and reward regulation were context specific, we fit the model to behavioral data for each trial type separately. To do so, we ordered t_p values associated with the same trial type from each behavioral session chronologically and treated them as consecutive samples from the model irrespective of the actual trial lag between them. Although this strategy made the model fitting more tractable, the inferred length constants in units of trials are likely smaller than the true values in the data. Methods for fitting behavioral data to the RSGP model were adapted from *Gaussian Processes for Machine Learning* (Rasmussen and Williams, 2006). The objective was to maximize the marginal likelihood of the observed data with respect to the hyperparameters $\{l_{SE}, \sigma_{SE}, l_{RS}, \sigma_{RS}, \sigma_0\}$. Using simulations, we found that optimization through searching the entire parameter space was inefficient and hindered convergence. Therefore, we implemented a two-step optimization. We first used the unrewarded trials to estimate l_{SE} and σ^2_{SE}, and then used those fits to search for the best fit of the remaining hyperparameters (l_{RS}, σ_{RS}, and σ_0) using all trials. The optimization of the multivariate likelihood function was achieved by line searching with quadratic and cubic polynomial approximations. The conjugate gradients was used to compute the search directions (Rasmussen and Williams, 2006). The landscape of likelihood indicated that the optimization was convex for a wide range of initial values (Figure S6A, Table S1).

The RSGP model fit to data provides a prediction of the distribution of t_p on each trial based on previous trials (t_p and reward history). We used this distribution to generate simulated values of t_p for each session, and repeated this process (n = 100) to estimate the distribution of $\mu(e^n)$ and $\sigma(e^n)$ in relation to e^{n-1} using the same analysis we applied to behavioral data (Figure 4C). To derive an estimate of the slow part of error, e_{slow}, we first fitted the RSGP to the behavior, and then used the reduced RSGP that only included the slow kernel ($\boldsymbol{K}_{SE}$) to predict the expected value of e_{slow}, i.e. the mean of a GP process governed by $\boldsymbol{K}_{SE}$.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Relationship between neural activity and the slow component of behavior

We used linear regression to examine whether and to what extent the population neural activity could predict the slow component of error (e_{slow}) inferred from the RSGP model fits to behavior (as described in previous section) using the following regression model:

$$e_{\text{slow}} = \boldsymbol{r\beta} + \beta_0$$

where $\boldsymbol{r}$ represents a matrix (nxN) containing spike counts within a 250 ms of N simultaneously recorded neurons across n trials, β_0 is a constant, and $\boldsymbol{\beta}$ is a N dimensional vector specifying the contribution of each neuron to e_{slow}. We used a random half of trials (training dataset) to find $\boldsymbol{\beta}$ and β_0 and the other half (validation dataset) to test the model, and quantified the success of the model by computing the Pearson correlation coefficient (Figure 5B, 5C, and 5D) between the e_{slow} inferred from RSGP model fits to behavior and e_{slow} predicted from the neural data using the regression model.

We initially tested the regression model using spike counts immediately before Set and later extended the analysis to different time points throughout the trial. To do so, we aligned spike times to various events throughout the trial (Cue, Tar, Set, Go) and tested the regression model every 125 ms around each event (4 time points after Cue, 3 time points before Tar, 3 time points after Tar, 4 time points before Set, 4 time points after Set and 4 time points before Go).

Statistical Analysis

Mean ± standard error of the mean (SEM) or median ± median absolute deviation (MAD) were used to report statistics. We detailed all the statistics in the Results and figure captions. All hypotheses were tested at a significance level of 0.01 and P-values were reported. We used t-tests to perform statistical tests on the following variables: (1) weber fraction (ration of standard deviation to mean of t_p), (2) cross correlation between pairs of trials of differents lag and trial type, (3) variance terms in the RSGP model (σ^2_{SE}, σ^2_{RS}, and σ^2_0) which were assumed to be normally distributed. We used one-tailed paired, two-tailed paired or two-sample *t*-tests depending on the nature of data and question. The length scale parameters of the RSGP model (l_{SE} and l_{RS}) were not normally distributed. Therefore, we used a one-way ANOVA to test whether the two were significantly different. We used a two-sample *F*-test to compare variability of production interval for different pair-trial condition (H0: equal variance).

Data Availability statement and Accession Code Availability Statements

Data that support the findings of this study are available from the corresponding author upon reasonable request. Matlab codes for simulating the RSGP are available at https://github.com/wangjing0/RSGP

Mathematical notation

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Symbol	Description
t_p^n	Production time in the n^th trial
e^n	$(t_p^n - t_t)/t_t$, relative error of production time in the n^th trial
$\mu(e)$	Mean of relative error
$\sigma(e)$	Standard deviation of relative error
$K(i,j)$	Covariance between trial i and trial j
σ^2	Signal variance or noise variance associated with a Gaussian process
l	Length scale of the squared-exponential covariance function
$\mathbf{r}^n(t)$	Population spiking at time t in the n^th trial
β	Regression coefficient
Z	Projection of population spiking activity onto a low dimensional representation

References

Afshar, A., Santhanam, G., Yu, B.M., Ryu, S.I., Sahani, M., and Shenoy, K.V. (2011). Single-trial neural correlates of arm movement preparation. Neuron *71*, 555–564.

Ajemian, R., D'Ausilio, A., Moorman, H., and Bizzi, E. (2013). A theory for how sensorimotor skills are learned and retained in noisy and nonstationary neural circuits. Proc. Natl. Acad. Sci. U. S. A. *110*, E5078–E5087.

Ames, K.C., Ryu, S.I., and Shenoy, K.V. (2014). Neural dynamics of reaching following incorrect or absent motor preparation. Neuron *81*, 438–451.

Ashmore, R.C., and Sommer, M.A. (2013). Delay activity of saccade-related neurons in the caudal dentate nucleus of the macaque cerebellum. J. Neurophysiol. *109*, 2129–2144.

Berman, R.A., and Wurtz, R.H. (2011). Signals conveyed in the pulvinar pathway from superior colliculus to cortical area MT. J. Neurosci. *31*, 373–384.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., and Ljung, G.M. (2015). Time Series Analysis: Forecasting and Control (John Wiley & Sons).

Carpenter, R.H., and Williams, M.L. (1995). Neural computation of log likelihood in control of

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

saccadic eye movements. Nature *377*, 59–62.

Chaisanguanthum, K.S., Shen, H.H., and Sabes, P.N. (2014). Motor variability arises from a slow random walk in neural state. J. Neurosci. *34*, 12071–12080.

Chen, X., Mohr, K., and Galea, J.M. (2017). Predicting explorative motor learning using decision-making and motor noise. PLoS Comput. Biol. *13*, e1005503.

Chen, Y., Ding, M., and Kelso, J.A.S. (1997). Long memory processes (1/f α type) in human coordination. Phys. Rev. Lett. *79*, 4501.

Church, R.M., and Broadbent, H.A. (1990). Alternative representations of time, number, and rate. Cognition *37*, 55–81.

Churchland, M.M., Afshar, A., and Shenoy, K.V. (2006). A central source of movement variability. Neuron *52*, 1085–1096.

Churchland, M.M., Yu, B.M., Cunningham, J.P., Sugrue, L.P., Cohen, M.R., Corrado, G.S., Newsome, W.T., Clark, A.M., Hosseini, P., Scott, B.B., et al. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nat. Neurosci. *13*, 369–378.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56.

Cohen, M.R., and Maunsell, J.H.R. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nat. Neurosci. *12*, 1594–1600.

Crossman, E.R.F.W. (1959). A THEORY OF THE ACQUISITION OF SPEED-SKILL∗. Ergonomics *2*, 153–166.

Dam, G., Kording, K., and Wei, K. (2013). Credit Assignment during Movement Reinforcement Learning. PLoS One *8*, e55352.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879.

Dayan, P., and Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. *8*, 429–453.

Dhawale, A.K., Smith, M.A., and Ölveczky, B.P. (2017). The Role of Variability in Motor Learning. Annu. Rev. Neurosci.

Fee, M.S., and Goldberg, J.H. (2011). A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. Neuroscience *198*, 152–170.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat. Neurosci. *12*, 1062–1068.

Gallistel, C.R., and Gibbon, J. (2000). Time, rate, and conditioning. Psychol. Rev. *107*,

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

289–344.

Gao, P., Trautmann, E., Yu, B.M., Santhanam, G., Ryu, S., Shenoy, K., and Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. Psychol. Rev. *84*, 279.

Gibbon, J., Church, R.M., and Meck, W.H. (1984). Scalar Timing in Memory. Ann. N. Y. Acad. Sci. *423*, 52–77.

Gilden, D.L., Thornton, T., and Mallon, M.W. (1995). 1/f noise in human cognition. Science *267*, 1837–1839.

Grossberg, S., and Schmajuk, N.A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. Neural Netw. *2*, 79–102.

Guo, Z.V., Inagaki, H.K., Daie, K., Druckmann, S., Gerfen, C.R., and Svoboda, K. (2017). Maintenance of persistent activity in a frontal thalamocortical loop. Nature *545*, 181–186.

Halassa, M.M., Chen, Z., Wimmer, R.D., Brunetti, P.M., Zhao, S., Zikopoulos, B., Wang, F., Brown, E.N., and Wilson, M.A. (2014). State-dependent architecture of thalamic reticular subnetworks. Cell *158*, 808–821.

Harris, C.M., and Wolpert, D.M. (1998). Signal-dependent noise determines motor planning. Nature *394*, 780–784.

Harris, K.D., and Thiele, A. (2011). Cortical state and attention. Nat. Rev. Neurosci. *12*, 509–523.

Hauser, C.K., Zhu, D., Stanford, T.R., and Salinas, E. (2018). Motor selection dynamics in FEF explain the reaction time variance of saccades to single targets. Elife *7*.

Hayden, B.Y., Pearson, J.M., and Platt, M.L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. Nat. Neurosci. *14*, 933–939.

Herzfeld, D.J., Kojima, Y., Soetedjo, R., and Shadmehr, R. (2015). Encoding of action by the Purkinje cells of the cerebellum. Nature *526*, 439–442.

Hoshi, E., Tremblay, L., Féger, J., Carras, P.L., and Strick, P.L. (2005). The cerebellum communicates with the basal ganglia. Nat. Neurosci. *8*, 1491–1493.

Huang, C., Ruff, D.A., Pyle, R., Rosenbaum, R., Cohen, M.R., and Doiron, B. (2019). Circuit Models of Low-Dimensional Shared Variability in Cortical Networks. Neuron *101*, 337–348.e4.

Huang, V.S., Haith, A., Mazzoni, P., and Krakauer, J.W. (2011). Rethinking motor learning and savings in adaptation paradigms: model-free memory for successful actions combines with internal models. Neuron *70*, 787–801.

Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

architecture in the cat's visual cortex. J. Physiol. *160*, 106–154.

Huberdeau, D.M., Krakauer, J.W., and Haith, A.M. (2015). Dual-process decomposition in human sensorimotor adaptation. Curr. Opin. Neurobiol. *33*, 71–77.

Ito, M. (2002). Historical review of the significance of the cerebellum and the role of Purkinje cells in motor learning. Ann. N. Y. Acad. Sci. *978*, 273–288.

Izawa, J., and Shadmehr, R. (2011). Learning from sensory and reward prediction errors during motor adaptation. PLoS Comput. Biol. *7*, e1002012.

Jazayeri, M., and Shadlen, M.N. (2010). Temporal context calibrates interval timing. Nat. Neurosci. *13*, 1020–1026.

Jazayeri, M., and Shadlen, M.N. (2015). A Neural Mechanism for Sensing and Reproducing a Time Interval. Curr. Biol. *25*, 2599–2609.

Kaelbling, L.P., Littman, M.L., and Moore, A.W. (1996). Reinforcement Learning: A Survey. 1 *4*, 237–285.

Kao, M.H., Doupe, A.J., and Brainard, M.S. (2005). Contributions of an avian basal ganglia–forebrain circuit to real-time modulation of song. Nature *433*, 638.

Karlsson, M.P., Tervo, D.G.R., and Karpova, A.Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. Science *338*, 135–139.

Kato, H.K., Chu, M.W., Isaacson, J.S., and Komiyama, T. (2012). Dynamic sensory representations in the olfactory bulb: modulation by wakefulness and experience. Neuron *76*, 962–975.

Killeen, P.R., and Fetterman, J.G. (1988). A behavioral theory of timing. Psychol. Rev. *95*, 274–295.

Krakauer, J.W., and Mazzoni, P. (2011). Human sensorimotor learning: adaptation, skill, and beyond. Curr. Opin. Neurobiol. *21*, 636–644.

Kunimatsu, J., and Tanaka, M. (2016). Striatal dopamine modulates timing of self-initiated saccades. Neuroscience.

Kunimatsu, J., Suzuki, T.W., Ohmae, S., and Tanaka, M. (2018). Different contributions of preparatory activity in the basal ganglia and cerebellum for self-timing. Elife *7*.

Laming, D. (1979). Autocorrelation of choice-reaction times. Acta Psychol. *43*, 381–412.

Lara, A.H., Elsayed, G.F., Zimnik, A.J., Cunningham, J.P., and Churchland, M.M. (2018). Conservation of preparatory neural events in monkey motor cortex regardless of how movement is initiated. Elife *7*.

Lau, B., and Glimcher, P.W. (2007). Action and outcome encoding in the primate caudate nucleus. J. Neurosci. *27*, 14502–14514.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Lauwereyns, J., Watanabe, K., Coe, B., and Hikosaka, O. (2002). A neural correlate of response bias in monkey caudate nucleus. Nature *418*, 413–417.

Lee, S.-H., and Dan, Y. (2012). Neuromodulation of brain states. Neuron *76*, 209–222.

Lee, M.D., Zhang, S., Munro, M., and Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. Cogn. Syst. Res. *12*, 164–174.

Luck, S.J., Chelazzi, L., Hillyard, S.A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J. Neurophysiol. *77*, 24–42.

Machado, A. (1997). Learning the temporal dynamics of behavior. Psychol. Rev. *104*, 241–265.

Malapani, C., and Fairhurst, S. (2002). Scalar Timing in Animals and Humans. Learn. Motiv. *33*, 156–176.

Mauk, M.D., and Buonomano, D.V. (2004). The neural basis of temporal processing. Annu. Rev. Neurosci. *27*, 307–340.

Mcalonan, K., Cavanaugh, J., and Wurtz, R.H. (2008). Guarding the gateway to cortex with attention in visual thalamus. Nature *456*, 391–394.

Medina, J., and Lisberger, S. (2008). Links from complex spikes to local plasticity and motor learning in the cerebellum of awake-behaving monkeys. Nat. Neurosci. *11*, 1185–1192.

Merrill, W.J., Jr, and Bennett, C.A. (1956). The application of temporal correlation techniques in psychology. J. Appl. Psychol. *40*, 272.

Middleton, F.A., and Strick, P.L. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. Brain Cogn. *42*, 183–200.

Mitchell, J.F., Sundberg, K.A., and Reynolds, J.H. (2009). Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. Neuron *63*, 879–888.

Murakami, M., Shteingart, H., Loewenstein, Y., and Mainen, Z.F. (2017). Distinct Sources of Deterministic and Stochastic Components of Action Timing Decisions in Rodent Frontal Cortex. Neuron *94*, 908–919.e7.

Narain, D., Remington, E.D., Zeeuw, C.I.D., and Jazayeri, M. (2018). A cerebellar mechanism for learning prior distributions of time intervals. Nat. Commun. *9*, 469.

Narayanan, N.S., and Laubach, M. (2008). Neuronal correlates of post-error slowing in the rat dorsomedial prefrontal cortex. J. Neurophysiol. *100*, 520–525.

Ni, A.M., Ruff, D.A., Alberts, J.J., Symmonds, J., and Cohen, M.R. (2018). Learning and attention reveal a general relationship between population activity and behavior. Science *359*, 463–465.

Niell, C.M., and Stryker, M.P. (2010). Modulation of visual responses by behavioral state in

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

mouse visual cortex. Neuron *65*, 472–479.

Nikooyan, A.A., and Ahmed, A.A. (2015). Reward feedback accelerates motor learning. J. Neurophysiol. *113*, 633–646.

Olveczky, B.P., Andalman, A.S., and Fee, M.S. (2005). Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. PLoS Biol. *3*, e153.

Oprisan, S.A., and Buhusi, C.V. (2014). What is all the noise about in interval timing? Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20120459.

Paton, J.J., and Buonomano, D.V. (2018). The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. Neuron *98*, 687–705.

Pekny, S.E., Izawa, J., and Shadmehr, R. (2015). Reward-dependent modulation of movement variability. J. Neurosci. *35*, 4015–4024.

Rasmussen, C.E., and Williams, C.K.I. (2006). Gaussian process for machine learning (MIT press).

Remington, E.D., Narain, D., Hosseini, E.A., and Jazayeri, M. (2018a). Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. Neuron *98*, 1005–1019.e5.

Remington, E.D., Egger, S.W., Narain, D., Wang, J., and Jazayeri, M. (2018b). A Dynamical Systems Perspective on Flexible Motor Timing. Trends Cogn. Sci. *22*, 938–952.

Ruff, D.A., and Cohen, M.R. (2014). Attention can either increase or decrease spike count correlations in visual cortex. Nat. Neurosci. *17*, 1591–1597.

Saalmann, Y.B., Pinsk, M.A., Wang, L., Li, X., and Kastner, S. (2012). The pulvinar regulates information transmission between cortical areas based on attention demands. Science *337*, 753–756.

Santos, F.J., Oliveira, R.F., Jin, X., and Costa, R.M. (2015). Corticostriatal dynamics encode the refinement of specific behavioral variability during skill learning. Elife *4*, e09423.

Schmitt, L.I., Wimmer, R.D., Nakajima, M., Happ, M., Mofakham, S., and Halassa, M.M. (2017). Thalamic amplification of cortical connectivity sustains attentional control. Nature *545*, 219–223.

Sheahan, H.R., Franklin, D.W., and Wolpert, D.M. (2016). Motor Planning, Not Execution, Separates Motor Memories. Neuron *92*, 773–779.

Shmuelof, L., Huang, V.S., Haith, A.M., Delnicki, R.J., Mazzoni, P., and Krakauer, J.W. (2012). Overcoming Motor "Forgetting" Through Reinforcement Of Learned Actions. J. Neurosci. *32*, 14617–14621a.

Simen, P., Balci, F., de Souza, L., Cohen, J.D., and Holmes, P. (2011). A model of interval timing by neural integration. J. Neurosci. *31*, 9238–9253.

Smith, M.A., Ghazizadeh, A., and Shadmehr, R. (2006). Interacting adaptive processes with

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

different timescales underlie short-term motor learning. PLoS Biol. *4*, e179.

Sommer, M.A., and Wurtz, R.H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. Nature *444*, 374–377.

Staddon, J.E., and Higa, J.J. (1999). Time and memory: towards a pacemaker-free theory of interval timing. J. Exp. Anal. Behav. *71*, 215–251.

Sternad, D., and Abe, M.O. (2010). Variability, Noise, and Sensitivity to Error in Learning a Motor Task. In Motor Control, (Oxford University Press),.

Sul, J.H., Kim, H., Huh, N., Lee, D., and Jung, M.W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron *66*, 449–460.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (MIT Press).

Takikawa, Y., Kawagoe, R., and Hikosaka, O. (2002). Reward-dependent spatial selectivity of anticipatory activity in monkey caudate neurons. J. Neurophysiol. *87*, 508–515.

Thoroughman, K.A., and Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. Nature *407*, 742–747.

Tumer, E.C., and Brainard, M.S. (2007). Performance variability enables adaptive plasticity of “crystallized” adult birdsong. Nature *450*, 1240–1244.

Verstynen, T., and Sabes, P.N. (2011). How each movement changes the next: an experimental and theoretical study of fast adaptive priors in reaching. J. Neurosci. *31*, 10050–10059.

Vinck, M., Batista-Brito, R., Knoblich, U., and Cardin, J.A. (2015). Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. Neuron *86*, 740–754.

Vyas, S., Even-Chen, N., Stavisky, S.D., Ryu, S.I., Nuyujukian, P., and Shenoy, K.V. (2018). Neural Population Dynamics Underlying Motor Learning Transfer. Neuron *97*, 1177–1186.e3.

Wagenmakers, E.-J., Farrell, S., and Ratcliff, R. (2004). Estimation and interpretation of 1/falpha noise in human cognition. Psychon. Bull. Rev. *11*, 579–615.

Wang, J., Narain, D., Hosseini, E.A., and Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. Nat. Neurosci. *21*, 102–110.

Weiss, B., Coleman, P.D., and Green, R.F. (1955). A stochastic model for time-ordered dependencies in continuous scale repetitive judgments. J. Exp. Psychol. *50*, 237.

Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A., and Cohen, J.D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. J. Exp. Psychol. Gen. *143*, 2074–2081.

Wimmer, R.D., Schmitt, L.I., Davidson, T.J., Nakajima, M., Deisseroth, K., and Halassa, M.M. (2015). Thalamic control of sensory selection in divided attention. Nature *526*, 705–709.

Wolpert, D.M., Diedrichsen, J., and Flanagan, J.R. (2011). Principles of sensorimotor learning.

bioRxiv preprint first posted online Mar. 28, 2019; doi: http://dx.doi.org/10.1101/583328. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Nat. Rev. Neurosci. *12*, 739–751.

Wu, H.G., Miyamoto, Y.R., Gonzalez Castro, L.N., Ölveczky, B.P., and Smith, M.A. (2014). Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. Nat. Neurosci. *17*, 312–321.

Xiong, Q., Znamenskiy, P., and Zador, A.M. (2015). Selective corticostriatal plasticity during acquisition of an auditory discrimination task. Nature.

Yasuda, M., and Hikosaka, O. (2015). Functional territories in primate substantia nigra pars reticulata separately signaling stable and flexible values. J. Neurophysiol. *113*, 1681–1696.

Zhou, H., Schafer, R.J., and Desimone, R. (2016). Pulvinar-Cortex Interactions in Vision and Attention. Neuron *89*, 209–220.