1 Reinforcement regulates timing variability in thalamus

2
3 Jing Wang^{1,2,*}, Eghbal Hosseini², Nicolas Meirhaeghe³, Adam Akkad², Mehrdad Jazayeri^{2,*†}

4 ¹Department of Bioengineering, University of Missouri, Columbia, MO 65211 ²McGovern Institute for Brain
5 Research, Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge,
6 Massachusetts 02139 ³Harvard-MIT Division of Health Sciences & Technology, Cambridge, Massachusetts
7 02139, USA

8
9
10 ***Corresponding author statement:**
11 Correspondence may be directed to MJ (mjaz@mit.edu) or JW (jingwang.physics@gmail.com)

12
13 **†Lead contact:**
14 Mehrdad Jazayeri, Ph.D.
15 Associate Professor, Department of Brain and Cognitive Sciences
16 Investigator, McGovern Institute for Brain Research
17 Email: mjaz@mit.edu

18 **Abstract**

19 Learning reduces variability but variability can facilitate learning. This paradoxical relationship has made it
 20 challenging to tease apart sources of variability that degrade performance from those that improve it. We
 21 tackled this question in a context-dependent timing task requiring humans and monkeys to flexibly produce
 22 different time intervals with different effectors. Subjects' timing variability featured two novel and context-
 23 specific sources of variability: (1) slow memory-contingent fluctuations of the mean that degraded
 24 performance, and (2) fast reinforcement-dependent regulation of variance that improved performance.
 25 Signatures of these processes were evident across populations of neurons in multiple nodes of the cortico-
 26 basal ganglia circuits. However, only in a region of the thalamus involved in flexible control of timing were
 27 the slow performance-degrading fluctuations aligned to performance-optimizing regulation of variance.
 28 These findings provide direct evidence that the nervous system makes strategic use of exploratory variance
 29 to guard against other undesirable sources of variability.

30 Introduction

31 While interacting with a dynamic and uncertain environment, humans and animals routinely rely on learning
 32 through trial and error to optimize their behavior. This type of learning is most rigorously formulated within
 33 the framework of reinforcement learning (RL). The key idea behind RL is to compute a value function for all
 34 possible actions based on previous outcomes, and use those action-outcome relationships to guide future
 35 actions (Sutton and Barto, 1998). RL has been extremely valuable in explaining behavior when an agent
 36 has to choose among a small set of discrete options, such as classic multi-armed bandit problems (Daw et
 37 al., 2005). However, learning through trial and error is also critical for other behavioral settings where the
 38 action space is large and continuous (Dhawale et al., 2017). For example, this strategy seems particularly
 39 relevant when the agent has to adjust its movements based on a binary or scalar (unsigned) feedback
 40 (Chen et al., 2017; Dam et al., 2013; Izawa and Shadmehr, 2011; Nikooyan and Ahmed, 2015; Pekny et
 41 al., 2015; Shmuelof et al., 2012a; Wu et al., 2014). In this case, classic RL is not feasible as it requires the
 42 agent to compute the value function over an infinity of movement parameters (e.g., all the kinematic degrees
 43 of freedom associated with a simple reach and grasp movement). Moreover, in a continuous state space,
 44 the slightest source of variability such as the inherent noise in the motor system would degrade the agent's
 45 ability to update action-outcome relationships.

46 This raises the question of how humans and animals employ trial and error learning when the space of
 47 possibilities is continuous. One intriguing hypothesis is that the brain directly regulates variability to facilitate
 48 learning (Huang et al., 2011; Kao et al., 2005; Olveczky et al., 2005; Tumer and Brainard, 2007). The basic
 49 idea is to reduce variability when a trial yields reward (i.e., exploit previously rewarded action), and increase
 50 variability in the absence of reward (i.e., promote exploration of new possibilities). This strategy captures
 51 the spirit of RL in moderating exploration versus exploitation without explicitly computing a continuous value
 52 function. Several indirect lines of evidence support this hypothesis (Dhawale et al. 2019; Shmuelof et al.
 53 2012; Vaswani et al. 2015; Cashaback et al. 2017; Palidis et al. ; van der Kooij and Smeets 2019;
 54 Cashaback et al. 2019; Chen et al. 2017; Dam et al. 2013; Izawa and Shadmehr 2011; Nikooyan and
 55 Ahmed 2015; Pekny et al. 2015; Shmuelof et al. 2012; Wu et al. 2014). For example, humans learn more
 56 efficiently if the structure of their natural movement variability aligns with the underlying learning objective
 57 (Wu et al., 2014), reaching movements become more variable during periods of low success rate (Izawa
 58 and Shadmehr, 2011; Pekny et al., 2015), and metrics of saccadic eye movements become more variable
 59 in the absence of reward (Takikawa et al., 2002).

60 However, the relationship between variability and learning is nuanced. While increasing motor variability
 61 may play a direct role in the induction of learning (Dhawale et al., 2017), one of the primary functions of
 62 motor learning is to reduce variability (Crossman, 1959; Harris and Wolpert, 1998; Smith et al., 2006;
 63 Sternad and Abe, 2010; Thoroughman and Shadmehr, 2000; Verstynen and Sabes, 2011). This two-sided
 64 relationship between learning and variability has made it challenging to tease apart behavioral and neural
 65 signatures of variability that degrade performance from those that improve it. A rigorous assessment of the
 66 interaction between learning and variability demands two important developments. First, we need a method
 67 capable of teasing apart sources of variability that hinder performance from those that facilitate learning.
 68 Second, we need to verify that the underlying neural circuits rely on reinforcement to regulate the variability
 69 along task-relevant dimensions. To address these problems, we focused on a motor timing task in which
 70 monkeys produced different time intervals using different effectors. We first verified the presence of motor
 71 variability due to memory drift and reward-dependent adjustment in motor timing. We then developed a

72 generative model to explain how reward regulates variability and facilitates learning. Finally, we probed the
 73 underlying neural circuits and asked whether reward could regulate task-relevant variability within the
 74 nervous system. We recorded from multiple nodes of the cortico-basal ganglia circuits that were previously
 75 found to support monkeys' timing behavior (Wang et al., 2018). Results indicated that the variability across
 76 the population of thalamic neurons with projections to the dorsomedial frontal cortex (DMFC) was indeed
 77 regulated by reward along task-relevant dimensions in a context-specific manner.

78 Results

79 Cue-Set-Go task

80 Two monkeys were trained to perform a Cue-Set-Go (CSG) motor timing task (Figure 1A). On each trial, the
 81 animal aimed to produce either an 800 ms (Short) or a 1500 ms (Long) time interval (t_t) either with a saccade
 82 (Eye) or with a button press (Hand). Each trial was initiated by presenting a fixation cue ("Cue") at the center
 83 of the screen. The Cue consisted of a circle and a square. A combination of color and shape indicated the
 84 trial type. The four trial types, Eye-Short (ES), Eye-Long (EL), Hand-Short (HS), and Hand-Long (HL) were
 85 randomly interleaved throughout the session. After a random delay, a visual stimulus ("Tar") was flashed to
 86 the left or right of the screen. This stimulus specified the position of the saccadic target for the Eye trials and
 87 served no function in the Hand trials. After another random delay, the presentation of a "Set" flash around
 88 the fixation spot indicated the start of the timing period. Animals had to proactively initiate a saccade or a
 89 button press aiming at the desired t_t . We refer to the movement initiation as the "Go". We measured the
 90 produced interval, t_p , as the interval between Set and Go. To receive reward, the relative error, defined as e
 91 $= (t_p - t_t)/t_t$, had to be within a window (Figure 1A). On rewarded trials, the magnitude of reward decreased
 92 linearly with the size of the error. The width of the reward window was controlled independently for each trial
 93 type and was adjusted adaptively on a trial-by-trial basis (see Methods).

94 Animals learned to use the Cue and flexibly switched between the four trial types (Figure 1B). For both
 95 effectors, a robust feature of the behavior was that produced intervals (t_p) were more variable for the Long
 96 compared to Short (Figure 1C). This is consistent with the common observation that timing variability scales
 97 with the interval being timed (Gibbon, 1977; Malapani and Fairhurst, 2002). This *scalar variability* was
 98 manifest in the linear relationship between mean (μ) and standard deviation (σ) of t_p in each behavioral
 99 session (Figure 1C, the slope of the linear test was significantly positive, one-tailed t-test, $p < 0.001$, for
 100 monkey A, $df = 128$, $t = 32.12$; for monkey D, $p < 0.001$, $df = 163$, $t = 24.06$).

101

Figure 1. Task, behavior, and reward-dependency of variability. (A) The Cue-Set-Go task. (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 ($\sigma(t_p) = \beta_1 \mu(t_p)$). Regression slopes were positive and significantly different from zero for both effectors ($\beta = 0.087 \pm 0.02$ mean $\pm$ std for Eye and 0.096 ± 0.021 for Hand in Monkey A; $\beta_1 = 0.10 \pm 0.02$ 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 from the full distribution (Figure S1). The larger circles with black outline correspond to grand $\sigma(t_p)$ and $\mu(t_p)$ computed from combining t_p values for each trial type across all 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. The fraction rewarded was the ratio between the number of rewarded trials and total number of trials across all behavioral sessions

119 Deconstructing motor timing variability

120 The neurobiological basis of *scalar variability* is not understood. Models of interval timing have attributed
 121 scalar variability to a variety of processes including a variable clock, a noisy accumulation process, noisy
 122 oscillations, and errors related to storage and/or retrieval of an interval (Church and Broadbent, 1990; Gibbon
 123 et al., 1984; Grossberg and Schmajuk, 1989; Jazayeri and Shadlen, 2010; Killeen and Fetterman, 1988;
 124 Machado, 1997; Oprisan and Buhusi, 2014; Simen et al., 2011; Staddon and Higa, 1999). These models
 125 assume that variability is stationary and therefore, predict that $\sigma(t_p)$ has a fixed linear relationship to $\mu(t_p)$. To
 126 test this assumption, we analyzed local estimates of $\sigma(t_p)$ as a function of $\mu(t_p)$ across trials of the same type
 127 from blocks of 50 consecutive trials. The relationship between $\mu(t_p)$ and $\sigma(t_p)$ across blocks of trials was
 128 strikingly different from the predictions of scalar variability (Figure 1D, S1): local estimates of $\sigma(t_p)$ decreased
 129 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
 130 was smaller when the animal received more reward and vice versa (Figure 1D). This inverse relationship
 131 was readily evident from the strong negative correlation between $\sigma(t_p)$ and the fraction of reward ($p < 0.001$,
 132 $r = -0.48$ for Monkey A and $r = -0.50$ for Monkey D). Although this result rejects models of interval timing that
 133 assume stationarity, the fact that animals' behavior is modulated by reward is unsurprising. In our
 134 experiment, the only information animals could rely on to calibrate their behavior was the amount of reward.
 135 The relationship between reward and variability suggests that animals calibrate their timing behavior through
 136 ongoing adjustment of behavioral variability.

137 The non-stationarity of behavior was also evident from serial correlations of t_p between trials of the same
 138 type (e.g., all trials of ES within a session). In all four trial types, t_p exhibited significant serial correlations up
 139 to a trial lag of 20 or more ($p < 0.01$, dash lines: 1% and 99% confidence bounds by estimating the null
 140 distribution from shuffled series, Figure 2A). These fluctuations have been reported in many tasks (Gilden
 141 et al., 1995; Merrill and Bennett, 1956; Weiss et al., 1955) and can be relatively strong in movements
 142 (Chaisanguanthum et al., 2014), reaction times (Laming, 1979), and interval timing (Chen et al., 1997;
 143 Murakami et al., 2017). Importantly, these correlations were context-specific; i.e. serial correlations were
 144 stronger between trials with the same effector and same t_t compared to trials that were associated with
 145 different t_t and/or different effector (see Figure 2B for statistics and Figure S2 for serial correlations between
 146 all trial types and saccade directions). The context-specificity of these correlations suggests that they were
 147 not solely due to non-specific fluctuations of internal states such as the overall level of alertness, which
 148 should persist across the four randomly interleaved trial types.

149 Since performance in CSG depends on an accurate memory of t_t , we hypothesized that these fluctuations
 150 may be due to slow drifts in memory. To test this hypothesis, we reasoned that the fluctuations should be
 151 smaller if the demands on memory were reduced. Therefore, we trained two monkeys, not previously
 152 exposed to the CSG task, to perform a control task in which t_t was measured on every trial, thereby
 153 minimizing the need to rely on memory (see Methods). As expected, t_p serial correlations diminished
 154 significantly in the control task (Figure S3; $p < 0.001$, paired sample t -test, $t_{19} = 4.9$, on the partial correlation
 155 between CSG and controls, trial lag was limited to 20 trials), suggesting that the slow fluctuations of t_p in the
 156 CSG task were a signature of drift in memory. For the remainder of our analyses, we will refer to these
 157 fluctuations as memory drifts. Note that this drift reflects a random process with no explicit baseline, and is
 158 consistent with a form of memory decay that is characterized by an increasingly broadening of the t_p
 159 distribution.

160

161

Figure 2. Context-dependent slow fluctuations of timing intervals. (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 a function of trial lag. Each curve shows the average partial correlations across all behavioral sessions. Four types of trials are shown in different colors. Filled circles: correlation values that are larger than the 1% and 99% confidence interval (dashed line). **(B)** Examples showing the Pearson correlation coefficients of t_p 's as a function of trial lag. HL-HL indicates the correlation was averaged across trials transitioning between HL and HL; 1% and 99% confidence intervals were estimated from the null distribution. Serial correlations were stronger between trials with the same effector and interval compared to trials with the same effector but different interval (***) dark gray, 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: $p < 0.001$, $n = 80$, $t_{79} = 5.8$), and trials of the different effector but same interval (***) light gray, monkey A: $p < 0.001$, $n = 80$, $t_{79} = 6.7$; monkey D: $p < 0.001$, $n = 80$, $t_{79} = 17.3$). See Figure S2 for transitions between other conditions, and for a comparison of context-specificity with respect to saccade direction.

173 Reward regulates variability in a context-specific manner

174 Memory drifts, if left unchecked, could lead to large excursions away from t_t , and ruin performance. To
 175 maintain a reasonably accurate estimate of t_t , another process must counter the drift in memory. In CSG,
 176 the only cue that animals can rely on is the trial outcome. Rewarded trials may be used to reinforce the
 177 memory while the absence of reward may be construed as evidence that the memory is inaccurate and has
 178 to be adjusted. Therefore, we hypothesized that modulations of behavioral variability by reward (Figure 1D)
 179 is a signature of an explore-exploit strategy aimed at countering memory drifts.

180 To test this hypothesis more directly, we defined relative error by $e = (t_p - t_t)/t_t$, and analyzed error statistics
 181 on a given trial (e^n) as a function of error in the preceding trial (e^{n-1}). We reasoned that errors should have
 182 two properties. First, because of the slow memory drifts, errors across consecutive trials should covary; i.e.,
 183 $\mu(e^n)$ should increase monotonically with e^{n-1} (Figure 3A, top). Second, an explore-exploit strategy predicts
 184 that variability should increase after low or absent reward (Figure 3A, bottom); i.e., $\sigma(e^n)$ as a function of e^{n-1}
 185 should have a U-shaped profile (Figure 3A, middle). Indeed, animals' behavior exhibited both properties:
 186 $\mu(e^n)$ increased monotonically with e^{n-1} , and $\sigma(e^n)$ had a U-shaped profile (Figure 3B, Figure S4A). We further
 187 confirmed these observations statistically using linear regression for $\mu(e^n)$ and quadratic regression for $\sigma(e^n)$
 188 (Table 1, see Methods).

189 We performed a number of additional analyses to further scrutinize this finding. For example, we asked
 190 whether the modulation of $\sigma(e^n)$ could be due to differences across (as opposed to within) behavioral
 191 sessions. Applying the same analysis to different grouping of trials verified that the relationship between
 192 variability and reward was present within individual behavioral sessions (Figure S4B). Similarly, we asked
 193 whether the effect of reward on variability was binary (present/absent) or graded (larger variability for lower
 194 reward). More refined binning of trials based on the magnitude of reward provided evidence that the effect
 195 was graded: $\sigma(e^n)$ was significantly larger in the absence of reward in comparison to low reward, and
 196 significantly larger in the low compared to high reward (Figure S4C). These results provide further evidence
 197 in support of the relationship between reward and behavioral variability.

198 Our earlier analysis indicated that the slow memory drifts in the behavior were context-specific (Figure 2B).
 199 For the reward modulation of variability to be effective, it must be also context-specific. That is, an increase
 200 in variability following a large error should only be present if the subsequent trial is of the same type.
 201 Otherwise, reinforcement of one trial type, say ES, may incorrectly adjust variability in the following trials of
 202 another type, say HL, and would interfere with the logic of harnessing reward to calibrate the memory of t_t .
 203 To test this possibility, we analyzed error statistics between pairs of consecutive trials associated with
 204 different types (Figure 3B open circles, Figure S4A; Table 1). Results indicated that (1) correlations were
 205 strongest between pairs of trials of the same type (as expected from our previous partial correlation analysis
 206 in Figure 2A), and (2) the modulation of variability was most strongly and consistently present across trials
 207 of the same type. Together, these results provide strong evidence that animals used reward to regulate
 208 behavioral variability in the next trial in a context-dependent manner and in accordance with an explore-
 209 exploit strategy.

210

211

212 Reward-dependent context-specific regulation of variability in humans

213 To ensure that our conclusions were not limited to data collected from highly trained monkeys, we performed
 214 a similar experiment in human subjects. In human psychophysical experiment, t_i varied from session to
 215 session, and subjects had to constantly adjust their t_p by trial-and-error. Similar to monkeys, human behavior
 216 exhibited long-term serial correlations (Figure S6A), and these correlations were context (effector) specific
 217 (Figure S6B). We performed the same analysis as in monkeys to characterize the dependence of $\mu(e^n)$ and
 218 $\sigma(e^n)$ on e^{n-1} . Results were qualitatively similar: $\mu(e^n)$ increased monotonically with e^{n-1} verifying the presence
 219 of slow drifts, and $\sigma(e^n)$ had a U-shaped with respect to e^{n-1} indicating that subjects used the feedback to
 220 regulate their variability (Figure 3C and Table 1). Finally, similar to monkeys, the effect of reward on variability
 221 was context-specific (Figure 3C). This result suggests that the memory of a time interval is subject to slow
 222 drifts and humans and monkeys use reward-dependent regulation of variability as a general strategy to
 223 counter these drifts and improve performance.

224 **Figure 3. Rapid and context-dependent modulation of behavioral variability by reward.** (A) Illustration of the expected effect
 225 of serial correlations and reward-dependent variability. Top: Positive serial correlations between produced intervals (t_p) creates a
 226 positive correlation between consecutive errors, and predicts a monotonic relationship between the mean of error in trial n , $\mu(e^n)$,
 227 and the value of error in trial $n-1$ (e^{n-1}). Middle: Variability decreases with reward. This predicts a U-shaped relationship between the
 228 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
 229 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
 230 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. Bottom: the
 231 reward expectation as a function of e^{n-1} . The reward expectation was computed by averaging the magnitude of reward received
 232 across trials. In the same effector, variability increased significantly after an unrewarded trials compared to a rewarded trials
 233 (horizontal line, two-sample F-test for equal variances, *** $p < 0.001$) for both large positive errors (Monkey A: $F(11169,10512) =$
 234 1.09 , Monkey D: $F(18540,13478) = 1.76$) and large negative errors (Monkey A: $F(8771,9944) = 1.40$, Monkey D: $F(21773,14889) =$
 235 1.62). The variability after an unrewarded trial of the same effector was significantly larger than after an unrewarded trial of the other
 236 effector (vertical line, two-sample F-test for equal variances, *** $p < 0.001$) for both large positive errors (Monkey A: $F(11169,8670) =$
 237 1.20 , Monkey D: $F(18540,7969) = 1.32$) and large negative errors (Monkey A: $F(8771,5994) = 1.26$, Monkey D: $F(21773,7179) =$
 238 1.27). (C) Same as (B) for human subjects. In humans, the variability was also significantly larger after a negatively reinforced trial
 239 compared to a positively reinforced trial (horizontal line, two-sample F-test for equal variances, *** $p < 0.001$) for both large positive
 240 errors ($F(5536,5805) = 1.19$) and large negative errors ($F(9366,9444) = 1.11$). The variability after a positively reinforced trial of the
 241 same effector was significantly lower than after a positively reinforced trial of the other effector (vertical line, two-sample F-test for

242 equal variances, *** $p < 0.001$, $F(14497, 15250) = 1.10$). For humans, the reward expectation was defined as the ratio between the
243 number of trials with positive feedback and total number of trials.

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

244

245 **Table 1. Quantitative assessment of the dependence of $\mu(e^n)$ and $\sigma(e^n)$ on e^{n-1} .** We used linear regression ($\mu(e^n) = m_0 + m_1 e^{n-1}$)
246 to relate $\mu(e^n)$ to e^{n-1} , and quadratic regression ($\sigma(e^n) = s_0 + s_1 e^{n-1} + s_2 (e^{n-1})^2$) to relate $\sigma(e^n)$ to e^{n-1} . Fit parameters and the corresponding
247 confidence intervals [1%, 99%] are tabulated for each monkey and for humans. We compared the magnitude of fit parameters
248 between the same versus different effector conditions (bold: significantly different).

249 **Trial outcome causally impacts behavioral variability**

250 So far, our results establish a relationship between trial outcome and behavioral variability. However, since
 251 in our experiment error size and trial outcome were deterministically related (larger error led to reduced
 252 reward), it is not possible to determine which of these two factors had a causal impact on behavioral
 253 variability. More specifically, our current findings are consistent with two models. In one model, trial outcome
 254 regulates behavioral variability (Figure 4A; “Causal”). In another, variability is influenced by various upstream
 255 factors such as error magnitude, but not by trial outcome (Figure 4A; “Correlational”). To distinguish between
 256 these two models, we carried out a new human psychophysics experiment (Figure 4B) in which trial outcome
 257 was probabilistic (Figure 4C) so that the presence or absence of a “correct” feedback was not fully
 258 determined by the magnitude of the error. This design enabled us to analyze the effect of trial outcome
 259 (“correct” versus “incorrect”) on behavioral variability separately for small and large errors.

260 We compared behavioral variability after “correct” and “incorrect” trials separately for small (Figure 4D, left)
 261 and large (Figure 4D, right) errors. Across subjects and irrespective of the size of error, variability was
 262 significantly larger after incorrect compared to correct trials (Figure 4D). We also confirmed that the
 263 association between variability and trial outcome was not mediated by the size of error (Figure S5C). This
 264 result substantiates our hypothesis and rejects all models in which the variability is not explicitly dependent
 265 on the trial outcome.

Figure 4. Causal effect of reward on behavioral variability. (A) Alternative models of the relationship between the outcome of the preceding trial (Rew^{n-1}) and behavioral variability in the current trial (σ^n). Left: A model in which Rew^{n-1} causally controls σ^n . In this model, various factors (e.g., memory drift) may determine the size of error (e^{n-1}), e^{n-1} determines Rew^{n-1} , and Rew^{n-1} directly regulates σ^n . Right: A model in which the relationship between Rew^{n-1} and σ^n is correlational. In this model, Rew^{n-1} is determined by e^{n-1} , and σ^n may be controlled by various factors (including e^{n-1}) but not directly by Rew^{n-1} . (B,C) A time interval production task with probabilistic feedback to distinguish between the two models in (A). (B) Trial structure. The subject has to press the spacebar to initiate the trial. During the trial, the subject is asked to hold their gaze on a white fixation spot presented at the center of the screen. After a random delay, a visual ring ('Set') is flashed briefly around the fixation spot. The subject has to produce a time interval after Set using a delayed key press. After the keypress, the color of the fixation spot changes to red or green to provide the subject with feedback (green: "correct", red: "incorrect"). (C) Top: A schematic illustration of the distribution of relative error between the produced interval (t_p) and the target interval (t_t), computed as $(t_p - t_t)/t_t$. Bottom: After each trial, the feedback is determined probabilistically: the subject is given a "correct" feedback with the probability of 0.7 when t_p is within a window around t_t , and with the probability of 0.3 when errors are outside this window. The window length was adjusted based on the overall behavioral variability so that each subject receives approximately 50% "correct" feedback in each behavioral session. (D) The causal effect of the outcome of the preceding trial on behavioral variability in the current trial. Left: Scatter plot shows the behavioral variability after "incorrect" (ordinate) versus "correct" (abscissa) trials, for all five subjects, after trials in which e^{n-1} was small (inset). Results for the positive and negative errors are shown separately (with '+' and '-' symbols, respectively). Right: Same as the left panel for trials in which e^{n-1} was large (inset). In both panels, the variability across subjects was significantly larger following incorrect compared to correct trials ($p < 0.001$, paired sample t-test, $t_{199} = 12.8$ for small error, and $t_{199} = 13.7$ for large error, see Methods).

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

289 Our results so far indicate that errors in t_p are governed by two key factors: long-term serial correlations
 290 creating local and persistent biases in t_p , and short-term modulations of t_p variability by reward (see Figure
 291 S7 for an analysis of the time lag of the effect of reward). Intuitively, this could provide the means for an
 292 efficient control strategy: when bias due to memory drift increases, error increases, reward drops, and the
 293 animal seeks to find the correct target interval by increasing variability. Indeed, several recent behavioral
 294 experiments have found evidence that is qualitatively consistent with this control strategy (Cashaback et al.,
 295 2019; Chen et al., 2017; Dam et al., 2013; Izawa and Shadmehr, 2011; Nikooyan and Ahmed, 2015; Pekny
 296 et al., 2015; Shmuelof et al., 2012a; Wu et al., 2014). To assess this idea rigorously, we aimed to develop a
 297 generative model that could emulate this control strategy. Initially, we considered two model classes:
 298 autoregressive models that can readily capture serial correlations (Wagenmakers et al., 2004) and RL
 299 models that use reward to guide behavior (Kaelbling et al., 1996; Sutton and Barto, 1998). However, these
 300 two are generally incompatible: classic autoregressive models are insensitive to reward, and RL models
 301 cannot accommodate serial correlations that are unrelated to reward. Moreover, RL models are usually
 302 geared toward problems with discrete action spaces such as multi-armed bandit problems (Dayan and Daw,
 303 2008), and adaptation of RL to continuous variables using Markovian sampling strategies (Haith and
 304 Krakauer, 2014) cannot capture the observed behavioral characteristics (Figure S5A and B).

305 Based on these results, we concluded that the generative process must have two distinct components, one
 306 associated with long-term serial correlations due to the memory drift, and another, with the short-term effect
 307 of reward on variability. Accordingly, we sought to develop a Gaussian process (GP) model that could
 308 capture both effects. This choice was motivated by three factors: 1) GPs automatically describe observations
 309 up to their second order statistics, which are the relevant statistics in our data, 2) GPs offer a nonparametric
 310 Bayesian fit to long-term serial correlations, and 3) as we will describe, GPs can be readily augmented to
 311 implement short-term control of variability based on reward.

312 GPs are characterized by a covariance matrix – also known as the GP kernel – that specifies the degree of
 313 dependence between samples, and thus determines how slowly the samples fluctuate. The most common
 314 and mathematically convenient formulation of this kernel is known as the “squared exponential” kernel
 315 function, denoted K_{SE} (Figure 5A, top left):

$$316 \quad K_{SE}(n, n-r) = \exp\left(-\frac{r^2}{2l_{SE}^2}\right)$$

317 In K_{SE} , the covariance between any two samples (indexed by n and $n-r$) drops exponentially as a function of
 318 temporal distance (r) and the rate of drop is specified by the *characteristic length parameter*, l_{SE} . When l_{SE}
 319 is small, samples are relatively independent, and when it is large, samples exhibit long range serial
 320 correlations (Figure 5A, left, 2nd and 3rd row).

321 Using GPs as the backbone of our model, we developed a reward-sensitive GP (RSGP) whose kernel
 322 (K_{RSGP}) is the weighted sum of two kernels, a classic squared exponential kernel (K_{SE}) scaled by σ_{SE}^2 , and
 323 a reward-sensitive kernel (K_{RS}) scaled by σ_{RS}^2 (Figure 5A, top middle). A third diagonal matrix ($\sigma_o^2 I$) was also
 324 added to adjust for baseline variance:

$$K_{RSGP}(n, n - r) = \sigma_{SE}^2 K_{SE}(n, n - r) + \sigma_{RS}^2 K_{RS}(n, n - r) + \sigma_0^2 I$$

K_{RS} also has the form of squared exponential with a length parameter, l_{RS} . However, the covariance terms were non-zero only for rewarded trials (Figure 5A, top middle). The reward was a binary variable determined by an acceptance window around t_i . This allows rewarded samples to have higher leverage on future samples (i.e., higher covariance), and this effect drops exponentially for trials farther in the future.

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

Intuitively, RSGP operates as follows: K_{SE} captures the long-term covariance across samples (Figure 5A, 3rd row, left). K_{RS} regulates shorter-term covariances (Figure 5A, 3rd row, middle) and allows samples to covary more strongly with recent rewarded trials (Figure 5A, bottom, middle). Non-zero values in K_{RS} after rewarded trials strengthen the correlation between samples and lead to a reduction of variability, whereas zero terms after unrewarded trials reduce correlations and increase variability. Using simulations of the full model with K_{RSGP} as well as reduced models with only K_{SE} or K_{RS} (Figure 5A, 2nd row), we verified that both kernels were necessary and that the full RSGP model was capable of capturing the two key features (Figure 5A, 4th and 5th row). Moreover, we used simulations to verify that parameters of the model were identifiable; i.e., fits of the model parameters to simulated data robustly recovered the ground truth (Table S1).

We fitted the RSGP model to behavior of both monkeys and humans (Figure 5B, Figure S8). The recovered characteristic length associated with serial correlations (l_{SE}) were invariably larger than that of the reward-dependent kernel (l_{RS}) (Monkeys: $l_{SE} = 20.6 \pm 21.4$, $l_{RS} = 2.0 \pm 0.7$; Humans: $l_{SE} = 19.2 \pm 21.8$, $l_{RS} = 1.3 \pm 0.4$; Median $\pm$ MAD). The model fit of variances (σ_{SE} , σ_{RS} and σ_0) are shown in Figure S8C. 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, $df = 219$, $t = 1.1$ 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 not significantly different 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, 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 5C, and Methods). 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.

Two aspects of the RSGP model are noteworthy. First, K_{RS} was formulated to capture the effect of reward between trials of the same type, and not transitions between trials associated with different types. Consequently, σ_{RS} for each trial type reflects the effect of reward on variability within the same trial type. The addition of kernels for transitions between effectors was unnecessary because the effect of reward was effector-specific (Figure 3B, Figure S4A). Second, the RSGP was designed to account for the effects of drift and reinforcement regardless of the target interval and thus did not distinguish between the Long and Short conditions. However, the RSGP can be easily reformulated to additionally account for the scaling of variability observed between the Long and Short conditions (Figure S9, and Methods). The RSGP was also able to capture the modulation of behavioral variability in the probabilistic reward experiment (Fig. S5 C and D)

362
363

Figure 5. 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 GP models, one using K_{SE} 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 the 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 from 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 sampling of trials is shown as shaded area but it is small and difficult to visually discern.

378

379 **Slow fluctuations in the cortico-basal ganglia circuits**

380 Recently, we identified a cortico-basal ganglia circuit that plays a causal role in animals' performance during
381 a flexible timing behavior (Wang et al., 2018). A combination of population data analysis and modeling
382 revealed that animals control the movement initiation time by adjusting a tonic signal in the thalamus that
383 sets the speed at which activity in the dorsomedial frontal cortex (DMFC) and caudate evolve toward an
384 action-triggering state. Based on these results, we hypothesized that memory drifts in behavior may be
385 accompanied by drifts of neural activity in these areas.

386 To test this hypothesis, we recorded separately from populations of neurons in candidate regions of DMFC,
387 caudate and thalamus. Based on previous work suggesting the importance of initial conditions in movement
388 initiation time (Carpenter and Williams, 1995; Churchland et al., 2006; Hauser et al., 2018; Jazayeri and
389 Shadlen, 2015; Lara et al., 2018; Remington et al., 2018a), we asked whether neural activity near the time
390 of Set (i.e., onset of motor timing epoch) could be used to predict the slow fluctuations in error, which we
391 denote by e_{slow} (Figure 6A, Right, red line). To estimate e_{slow} on a trial-by-trial basis, we relied on the RSGP
392 model, which readily decomposed the t_p time series into a slow memory-dependent and a fast reward-
393 dependent component. We fitted the RSGP in a context-specific manner and then inferred the value of e_{slow}
394 for each trial in a session (see Methods).

395 We measured spike counts within a 250-ms window before Set (Figure 6A, Left), denoted $\mathbf{r}$, and formulated
396 a multi-dimensional linear regression model to examine the trial-by-trial relationship between $\mathbf{r}$ and e_{slow} . We
397 computed the regression weight, $\boldsymbol{\beta}$, that when multiplied by $\mathbf{r}$, would provide the best linear fit to e_{slow} .
398 Considering the vector of spike counts in each trial as a point in a coordinate system where each axis
399 corresponds to the activity of one neuron ("state space"), $\boldsymbol{\beta}$ can be viewed as the direction along which
400 modulations of spike counts most strongly correlate with memory drifts. Accordingly, we will refer to the
401 direction associated with $\boldsymbol{\beta}$ as the drift direction, and will denote the strength of activity along that direction
402 by z ($z = \mathbf{r}\boldsymbol{\beta}$, Figure 6A, Right, blue line).

403 To test the hypothesis that memory drifts were accompanied by drifts of neural activity in the thalamus,
404 DMFC and caudate, we used a cross-validated procedure (see Methods) to compute $\boldsymbol{\beta}$, derive a trial-by-trial
405 estimate of z , and measure the correlation between z and e_{slow} on a session-by-session basis. We found
406 strong evidence that neural activity in all three areas exhibited slow fluctuations in register with memory
407 drifts. In 91%, ~79%, and ~45% of sessions, the activity along the drift direction in the thalamus (z_{Th}), DMFC
408 (z_{DMFC}), and caudate (z_{Cd}), respectively, was significantly correlated with e_{slow} (Figure 6B-D, $p < 0.01$, null
409 hypothesis test by shuffling trials).

410

Figure 6. Representation of slow fluctuations of behavior in population activity. (A) The schematics 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 providing a matrix containing the spike counts of all neurons across all trials. Middle: The spike counts across trials and neurons were used as the regressor in a multi-dimensional linear regression model with weight vector, β , to predict the slow component of error (e_{slow}). Right: We fitted the RSGP to errors (black dots, e) and then inferred e_{slow} . The plot shows the neural prediction (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 relative to the anterior commissure in each animal (AP: anterior posterior; ML: mediolateral). Recorded thalamic neurons were from a region of the thalamus with monosynaptic connections to DMFC (Inset: antidromically activated spikes in the thalamus.) Bottom: Histogram of the correlation coefficients between e_{slow} inferred from the RSGP model and z^n_{Th} (projection of thalamic population activity on drift direction) across recording sessions. Note that some correlations are negative because of cross-validation Black bars correspond to the sessions in which the correlation was significantly positive ($**p < 0.01$; hypothesis test by shuffling trial orders). The average correlation across all sessions and the average of those 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.

427 Reward-dependent regulation of variability in the thalamus

428 Next, we analyzed the statistics of neurally-inferred drift (z) across pairs of consecutive trials using the same
 429 approach we applied to error statistics in behavioral data (Figure 3B). To do so, we extracted pairs of (e^{n-1} ,
 430 z^n) for consecutive trials of the same type and binned them depending on the value of e^{n-1} . Since all three
 431 areas carried a signal reflecting memory drifts, we expected $\mu(z^n)$ in all the areas to have a monotonic
 432 relationship with e^{n-1} . Results were consistent with this prediction as evidenced by the slope of a linear
 433 regression model relating $\mu(z^n)$ to e^{n-1} (filled circles in Figure 7B-D top, Table 2). Moreover, this relationship
 434 was absent for consecutive trials associated with different effectors (open circles in Figure 7B-D top, Table
 435 2). These results substantiate the presence of context-specific fluctuations of neural activity in register with
 436 drifts in animals' memory of t_i across populations of neurons in the thalamus, DMFC and caudate.

437 A critical question was whether, in any of these areas, reward regulates the variability of neural activity.
 438 Importantly, the reward-dependent modulation of variability should be restricted to the drift direction in the
 439 population activity (Figure 7A, bottom); otherwise, this strategy will not be able to effectively counter the
 440 degrading effect of memory drift in a context-dependent manner. In thalamus, $\sigma(z_{Th}^n)$ exhibited the
 441 characteristic U-shape profile with respect to e^{n-1} (Figure 7B bottom), which we verified quantitatively by
 442 comparing the variance of z_{Th}^n for rewarded and unrewarded trials ($p < 0.01$, two-sample F-test for equal
 443 variances on z_{Th}^n , $F(5608,4266) = 1.08$ for negative e^{n-1} and $p < 0.001$ for positive e^{n-1} , $F(4723,6125) =$
 444 1.28), and by testing the profile using quadratic regression (see Methods; Table 2).

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

451 Unlike the thalamus, in DMFC and caudate, variability along the memory drift was relatively independent of
 452 e^{n-1} (Figure 7C bottom, 6D bottom); i.e., $\sigma(z_{DMFC}^n)$ and $\sigma(z_{Cd}^n)$ were not significantly different after rewarded
 453 and unrewarded trials (two-sample F-test for equal variances, $F(6244,4818) = 0.87$, $p = 0.99$ for the negative
 454 e^{n-1} ; $F(4825,7572) = 1.002$, $p = 0.021$ for the positive e^{n-1} , see Figure S10 for each animal separately). We
 455 verified the absence of a U-shape profile using quadratic regression (see Methods; Table 2) and ensured
 456 that the lack of modulation in DMFC and caudate compared to the thalamus was not because of a difference
 457 in the number of simultaneously recorded neurons (Figure S11 B and C). Together, these results provide
 458 evidence that the effect of reward in DMFC and caudate was not aligned to the β associated with the slow
 459 fluctuations. We note, however, that an unconstrained decoding strategy aimed at simultaneously capturing
 460 both components of errors can find directions along which the slow fluctuations and the effect of reward are
 461 aligned in all three areas (Figure S12).

Figure 7. 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 dimensions (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_i). 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 the 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 and 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 (β) 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 in the same format as in Figure 3B (thick lines: same effector; thin lines: different effectors). (C) and (D) Same as (B) for population activity in DMFC and caudate. See Figure S10 for the result of individual animal.

Brain area	Parameters		Same effector	Different effector
Thalamus	$\mu(z^n)$	m_1	0.13 [0.090 0.18]	0.0038 [-0.025 0.032]
		m_0	0.008 [0.001 0.016]	0.0041 [-0.001 0.009]
	$\sigma(z^n)$	s_2	0.26 [0.19 0.34]	0.0031 [-0.031 0.037]
DMFC	$\mu(z^n)$	s_1	0.005 [-0.013 0.023]	-0.001 [-0.01 0.0067]
		s_0	0.06 [0.052 0.059]	0.058 [0.057 0.0597]
	$\sigma(z^n)$	s_2	0.071 [-0.003 0.12]	0.0051 [-0.037 0.047]
Caudate	$\mu(z^n)$	s_1	0.0085 [-0.0022 0.019]	-0.0004 [-0.01 0.009]
		s_0	0.046 [0.044 0.048]	0.0487 [0.047 0.051]
	$\sigma(z^n)$	s_2	0.068 [0.016 0.12]	0.007 [-0.016 0.030]
	$\mu(z^n)$	m_1	0.0068 [-0.002 0.015]	0.018 [0.014 0.022]
		m_0		
	$\sigma(z^n)$	s_2	0.069 [-0.005 0.14]	-0.002 [-0.064 0.060]
	$\mu(z^n)$	s_1	0.0006 [-0.009 0.011]	0.001 [-0.013 0.015]
		s_0	0.048 [0.046 0.050]	0.048 [0.045 0.051]
	$\sigma(z^n)$	s_2		

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., $\mu(z^n) = m_0 + m_1 e^{n-1}$. 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., $\sigma(z^n) = s_0 + s_1 e^{n-1} + s_2 (e^{n-1})^2$. 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).

Our analysis so far focused on spike counts in a fixed 250-ms time window before Set. However, the effect of reward on variability in the thalamus (absent in DMFC and caudate) was present throughout the trial, and extended as far back as the preceding feedback (Figure 8A top; *** $p < 0.001$, two-sample F-test for equal variances). Together, these results suggest that reward exerts its influence on behavior by regulating the variability of thalamic signals driving movement initiation time.

487

Figure 8. Reward-dependent neural variability throughout the trial. (A) Average standard deviation of population activity in the 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 6A, 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. Black square: the time window used for the rest of this paper. We grouped e^{n-1} in to 12 bins, 6 for negative e^{n-1} and 6 for positive e^{n-1} . Results are shown separately for conditions in which trials $n-1$ 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} ; H_0 : 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 two animals. Figure S13 shows the results of the same analysis for each animal separately.

500 Discussion

501 Variability is a ubiquitous property of behavior that can either degrade performance or promote learning
 502 through exploration. Here, we were able to advance our understanding of the function and neurobiology of
 503 variability in three directions. First, we found an important role for memory drifts in timing variability, and
 504 showed that humans and monkeys use reward to calibrate memory against such drifts. Second, using model-
 505 based analysis of behavior, we showed that reward promotes learning through adjustment of variability.
 506 Finally, we characterized the neural underpinnings of variability associated with memory drifts and
 507 reinforcement across populations of neurons within the cortical-basal ganglia circuits.

508 Role of memory drift in behavioral variability

509 A key feature of the behavioral data was the presence of slow fluctuations in produced intervals leading to
 510 serial correlations extending over minutes and across dozens of trials. These fluctuations have been reported
 511 in various behavioral tasks (Chen et al., 1997; Gilden et al., 1995; Murakami et al., 2017), and are typically
 512 attributed to fatigue, arousal or other nonspecific factors modulating internal states (Harris and Thiele, 2011;
 513 Kato et al., 2012; Lee and Dan, 2012; Luck et al., 1997; Niell and Stryker, 2010; Vinck et al., 2015). Although
 514 slow internal state modulations are likely to present in our experiment, they cannot be the only driver since
 515 the serial correlations were strongly context-specific. Based on these observations, we reasoned that these
 516 fluctuations may reflect drifts in memory. This interpretation was consistent with the results of our control
 517 experiment showing diminished serial correlations when memory demands were reduced (Figure S3).
 518 However, further work is needed to fully characterize the nature of these slow fluctuations. For example, the
 519 slow component may be in part a reflection of an active averaging process to maintain a stable memory
 520 (Joiner and Smith, 2008), as suggested by error-based motor learning studies (Huberdeau et al., 2015; Smith
 521 et al., 2006; Wolpert et al., 2011).

522 A more puzzling observation was the specificity of memory drifts with respect to the target interval for the
 523 same effector. We currently do not have a definite explanation for this result, but our previous work in the
 524 domain of time interval production (Wang et al., 2018) and reproduction (Remington et al., 2018a) as well
 525 as others studies in the motor system (Afshar et al., 2011; Ames et al., 2014; Hauser et al., 2018; Sheahan
 526 et al., 2016; Vyas et al., 2018) suggest that several aspects of movement control can be understood in terms
 527 of adjusting inputs and initial conditions of a dynamical system (Churchland et al., 2012; Remington et al.,
 528 2018b). Accordingly, the interval specificity of the memory drifts suggests that distinct patterns of neural
 529 activity set the interval-dependent input and/or initial condition, which is consistent with our previous work
 530 (Wang et al., 2018).

531 Role of reinforcement in behavioral variability

532 In our task, reinforcement is the only information provided experimentally that subjects can use to calibrate
 533 their memory of the target interval. Therefore, the computational demands in our task fall squarely within the
 534 framework of reinforcement learning (RL). Most existing RL models have focused on experimental settings
 535 in which the agent faces a discrete set of options and/or a limited action space (Daw et al., 2006; Hayden et
 536 al., 2011; Lee et al., 2011; Wilson et al., 2014). In these situations, RL models posit that the agent keeps
 537 track of the value of available options and adopts a suitable policy to choose among them (Sutton and Barto,
 538 1998). In our task, subjects have to use reinforcement to choose the correct interval, which is a continuous

variable. In theory, RL can be extended to continuous variables, but doing so would require the brain to represent an infinite-dimensional value function, which seems implausible. Moreover, in this case, noise associated with producing an interval would cause the produced interval to differ from what was intended, and that would interfere with the agent's ability to correctly evaluate the intended action.

Due to these complications, several recent studies have proposed an alternative explore-exploit strategy for continuous variables in which the agent uses reinforcement to directly regulate behavioral variability (Pekny et al. 2015; Santos et al. 2015; Wu et al. 2014; Dhawale et al. 2017; Fee and Goldberg 2011; Tumer and Brainard 2007). Our results were consistent with this view: variability was modulated by reward in a manner that was consistent with an explore-exploit strategy. This hypothesis was strongly supported by the results of our psychophysical experiment using a probabilistic reward schedule, which confirmed that the effect of trial outcome on behavioral variability was causal and independent of the magnitude of error (Figure 4). Another key observation was that the effect of reward on behavioral variability was context-specific (Figure 3); 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. Given that the memory drifts were also context-specific, this finding indicates that one important function of reward-dependent regulation of variability is to counter memory drifts.

Extending reinforcement learning to continuous variables

Reinforcement learning through reward-dependent modulation of behavioral variability is an appealing idea that is gaining prominence, especially within the domain of motor learning (Dhawale et al., 2017). Here, we developed a reward-sensitive Gaussian process (RSGP) model to rigorously characterize and quantify this learning process. RSGP was able to capture both the long-term effect of memory drift, and the short-term effect of reward on variability. It also captured behavioral observations in the probabilistic reward experiment where we validated the causal effect of reward on variability (Figure S5C, D). A recent motor learning study in rodents found that the effect of reward on behavioral variability may last a few trials into the future (Dhawale et al. 2019). This was also evident in our data (Figure S7), and RSGP provided a quantitative measure of the temporal extent of this effect (Figure S8). As such, we expect RSGP to help future studies quantify the strength and persistence with which reinforcement guards against ubiquitous nonstationarities in behavior (Gilden et al. 1995; Weiss et al. 1955; Merrill and Bennett 1956; Laming 1979; Chaisanguanthum et al. 2014; Murakami et al. 2017; Chen et al. 1997).

Memory drift in the cortico-basal ganglia circuits

Behavioral variability in timing tasks is thought to have a multitude of distributed biophysical and synaptic origins (Gibbon et al., 1984; Mauk and Buonomano, 2004; Paton and Buonomano, 2018). Several studies have been able to trace this variability *in-vivo* to spiking activity of neurons in cortico-basal ganglia circuits (Murakami et al. 2017; Murakami et al. 2014; Gouvêa et al. 2015; Merchant and Averbeck 2017; Wang et al. 2018; Dhawale et al. 2017). We previously found that a circuit comprised of DMFC, DMFC-projecting thalamus and caudate plays a causal role in the control of movement initiation time (Wang et al., 2018), and that certain patterns of activity in each area were correlated with behavioral variability on a trial-by-trial basis. Here, we additionally confirmed that population activity in these areas carries a signal correlated with slow fluctuations of behavior. 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 suggests that they may partially reflect context-specific memory drifts. The key feature of our task that enabled us to uncover this specificity was the alternation between different contexts on a trial-by-trial basis. Since many previous studies did not include 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 specific 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 of the underlying synaptic and biophysical drifts. It is likely that the memory has a distributed representation, in which case the drift may result from stochastic processes distributed across multiple brain areas. It is also possible that different contexts engage 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.

For different effectors, context-specificity may be attributed to different patterns of activity across the same population of neurons or by distinct populations of neurons. Our results in the cortico-basal ganglia circuits provided evidence for the former (Figure 7, S15). However, we cannot rule out the possibility that this context-specificity arises partially from execution noise within distinct effector-specific downstream areas (e.g., brainstem) as we did not record from those areas. However, execution noise is generally considered to be irreducible (Dhawale et al., 2017; Faisal et al., 2008), and is therefore not expected to be influenced by reinforcement. Previous work suggests that central noise (e.g., in the cortico-basal ganglia circuits) plays a significant role in output variability (van Beers, 2009; Churchland et al., 2006). It is this portion of variability that is likely subject to adjustments through reinforcement.

Reinforcement via thalamus

Next, we asked whether the variability of neural activity in the thalamus, DMFC and caudate was 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 the memory drifts. This hypothesis makes a specific prediction: 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 the thalamus but not in DMFC or caudate. In the thalamus, spike count variability along the drift direction increased after rewarded trials and decreased after unrewarded trials in a context-specific manner. Previous studies have reported that in the thalamus firing rates are modulated on a trial-by-trial basis by attention (Mcalonan et al., 2008; Saalman et al., 2012; Zhou et al., 2016) and rule/context-dependent computations (Schmitt et al., 2017; Wang et al., 2018). Our work demonstrates that modulation of thalamic activity may additionally subserve reward-based calibration of movement initiation times. It will be important for future studies to investigate whether this finding generalizes to other movement parameters.

The fact that the same effect was not present in DMFC and caudate serves as a negative control and thus strengthens our conclusions. However, this begs the question of why this regulation was not inherited by

DMFC and caudate, especially given that DMFC receives direct input from the region of the 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 (Berman and Wurtz, 2011; Hubel and Wiesel, 1962; 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.

Our approach for investigating whether the effect of reinforcement was aligned with memory drift was correlative: we used a cross-validated decoding strategy to infer the response pattern most strongly associated with memory drift and tested whether reinforcement exerted its effect along that pattern. According to this analysis, only in thalamus the two effects were appropriately aligned. However, we cannot rule out the possibility that behavior was controlled by a pattern of activity different from those we inferred from our analysis. Indeed, with a complementary analysis using an unconstrained decoding approach, we were able to find patterns of activity in all three brain areas that simultaneously reflected the effects of memory drift and reinforcement (Figure S12). Therefore, an important consideration for future work is to use perturbation methods to verify causally the direction in the state space associated with memory drifts and reinforcement.

However, other considerations are consistent with thalamus playing a strong role in controlling timing variability. For example, it has been shown that inactivation of thalamus has a much stronger effect on modulating animals' motor timing variability compared to DMFC and caudate (Wang et al., 2018). Moreover, it has been shown that the nature of signals in DMFC-projecting thalamus and DMFC during motor timing are indeed different: DMFC neurons have 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. This transformation may provide an explanation for why reward-dependent modulation of firing rates was evident in the 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 the estimation of instantaneous neural states is unreliable unless the number of recorded neurons 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 the 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 S16). In other words, our results agree with the interpretation that reward controls the variance of average firing rates in the thalamus, and this effect leads to the control of the variance of the speed at which neural trajectories evolve in DMFC and caudate.

One open question pertains to which 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). Modulation of variability prior to movement initiation has been reported in motor and premotor areas (Churchland et al., 2006, 2010), and distinct correlates of explore-exploit strategy has been found across multiple nodes of the frontal cortex (Ebitz et al., 2018; Hayden et al., 2011; Massi et al., 2018; Sarafyazd and Jazayeri, 2019; Tervo et al., 2014). Therefore, it is possible that modulation of variability in thalamus originates in the cortex. One salient example where cortical variability can be adjusted rapidly and in a behaviorally-relevant fashion is in the domain of attentional control where spatial cueing can lead 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 possible that a similar strategy is at play in the frontal cortex in the service of exploratory learning. However, to act as an effective learning mechanism, such 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 hypothesis that the brain uses reinforcement to regulate behavioral variability in a context-dependent manner.

686

687 **Materials and Methods**

688 Two adult monkeys (*Macaca mulatta*; one female, one male), and five human subjects (18-65 years, two
689 females and three males) participated in the main task. Two monkeys (both male) participated in the memory
690 control experiment. In addition, five more human subjects (18-65 years, two females and three males)
691 participated in the probabilistic reward task to test the causal effect of reward on variability. The Committee
692 of Animal Care and the Committee on the Use of Humans as Experimental Subjects at Massachusetts
693 Institute of Technology approved the animal and human experiments, respectively. All procedures
694 conformed to the guidelines of the National Institutes of Health.

695 **Animal experiments**

696 Monkeys were seated comfortably in a dark and quiet room. The MWorks software package
697 (<https://mworks.github.io>) running on a Mac Pro was used to deliver stimuli and to control behavioral
698 contingencies. Visual stimuli were presented on a 23 inch monitor (Acer H236HL, LCD) at a resolution of
699 1920x1080, and a refresh rate of 60Hz). Auditory stimuli were played from the computer's internal speaker.
700 Eye position was tracked with an infrared camera (Eyelink 1000; SR Research Ltd, Ontario, Canada) and
701 sampled at 1 kHz. A custom-made manual button, equipped with a trigger and a force sensor, was used to
702 register button presses.

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

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

724 The width of the reward acceptance window was adjusted adaptively on a trial-by-trial basis and
725 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 brief visual and auditory feedback. 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 two naive monkeys. In the control experiment, the animals 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. 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 the thalamus, 568 units in the 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 number of simultaneously recorded neurons across sessions were shown in Figure S11A (mean $\pm$ s.d.; Thalamus: 17.9 ± 9.0 ; DMFC: 9.2 ± 4.3 ; Caudate: 14.8 ± 6.5). The DMFC comprises supplementary eye field, dorsal supplementary motor area (i.e., excluding the medial bank), and pre-supplementary motor area. We recorded from the left hemisphere from Monkey A and right from Monkey D, which was contralateral to the preferred hand used in the button press. 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 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 μA) delivered to DMFC via low impedance tungsten microelectrodes (100 – 500 K Ω , Microprobes). A stainless steel 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 session lasted approximately 60 minutes. Each subject completed 2-3 sessions per week. Similar to monkeys, experiments were conducted using the MWorks. All stimuli were presented on a black background monitor. Subjects were instructed to hold their gaze on a fixation point and hold a custom made push button using their dexterous hand, 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. The 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_i 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.

A set of different human subjects performed the same timing interval production task, except that the nature of the feedback was provided probabilistically. Subjects received “correct” feedback with the probability of 0.7 when t_p was within a window around t_i , and with the probability of 0.3 when errors were outside that window (Fig. 4C). The window length was adjusted based on the overall behavioral variability so that each subject receives approximately 50% “correct” feedback in each behavioral session. Only one effector - hand pressing - was used in this causal experiment, as we have established that both the slow and reward regulated variability were effector specific. All the experimental parameters were set identical to the main experiment.

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), $N = 62$ sessions and $n = 59,297$ trials in humans, and $N = 51$ sessions and $n = 30,695$ trials in the probabilistic feedback experiment in human subjects. Behavioral data for the no-memory control task was collected in $N = 26$ sessions consisting of $n = 75,652$ trials in two naive monkeys ($N = 9$, $n = 32,041$ for monkey G and $N = 17$, $n = 43,611$ for monkey H).

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 unchanged 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 a successive autoregressive model with the 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 S6). 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.

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 S4 A and B). 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 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 S4A 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 trial types in two ways. First, for each effector, we combined the Short and Long trial types by normalizing t_p values by their respective t_i . The resulting variable was defined as relative error $e^n = (t_p^n - t_i)/t_i$. 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 seconds apart.

We examined the relationship between $\mu(e^n)$ and e^{n-1} using a linear regression model of the form $\mu(e^n) = m_0 + m_1 e^{n-1}$, and the relationship between $\sigma(e^n)$ and e^{n-1} using a quadratic regression model of the form $\sigma(e^n) = s_0 + s_1 e^{n-1} + s_2 (e^{n-1})^2$. We assessed the significance of the monotonic relationship between $\mu(e^n)$ and e^{n-1} by the slope of linear regression (m_1), and the significance of the U-shaped profile in $\sigma(e^n)$ by the square term of a quadratic regression (s_2). Note that the choice of these models was not theoretically motivated; we simply considered these regression models to be an approximate function for testing the profile of $\mu(e^n)$ and $\sigma(e^n)$ as a function of e^{n-1} . The 1% and 99% confidence intervals were defined as, e.g. $s_2 \pm t_{(.01, N-p)} SE(s_2)$, where s_2 is the estimated coefficient, $t_{(.01, N-p)}$ is the 1 percentile of t-distribution with N-p degrees of freedoms, and $SE(s_2)$ is the standard error of estimated s_2 .

We applied the same analysis to the probabilistic reward experiment in humans after grouping the trials based on whether the feedback in the previous trials was “correct” or “incorrect”. Before analyzing the variability, we removed outlier trials defined as trials in which t_p was more than three standard deviations away from the mean. Since subjects were allowed to initiate the trials, we also excluded pairs of adjacent trials that were more than 7 seconds apart. When combining data across sessions, we normalized the relative error across sessions and subjects. To avoid sampling bias, the same number of trials were drawn repeatedly after “correct” or “incorrect”. The standard error of the mean was computed from 100 repeats with replacement (Fig. 4D and S5C).

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

We constructed a reward-sensitive Gaussian process model whose covariance function, K_{RSGP} , is a weighted sum of two kernels, a traditional squared exponential kernel, for which we used subscript SE (K_{SE}), and a reward-sensitive kernel with subscript RS (K_{RS}). The two kernels contribute to K_{RSGP} through scale factors σ_{SE}^2 and σ_{RS}^2 , 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 rates of drop for K_{SE} and K_{RS} are specified by characteristic length parameters, l_{SE} and l_{RS} , respectively. The model also includes a static source of variance, $\sigma^2 oI$ (I stands for the identity matrix):

$$K_{RSGP}(n, n - r) = \sigma_{SE}^2 K_{SE}(n, n - r) + \sigma_{RS}^2 K_{RS}(n, n - r) + \sigma_0^2 I$$

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

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

Note that 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 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_{SE}^2 and σ_{RS}^2 by σ^2 and $(1 - \sigma^2)$, respectively where $\sigma^2 = 1.0, 0$, and 0.5 for the three examples (Figure 3A, Table S1).

for $i := 1 \dots n$ **do**

1. Given the previous value and reward history, infer the mean and The variance of t_p^n from the conditional distribution

$$[tp^1, tp^{n-1}, tp^n]^\top \sim \mathcal{N}\left(t_t, \begin{bmatrix} K_{(n-1) \times (n-1)}, K_{(n-1) \times 1} \\ K_{1 \times (n-1)}, K_{1 \times 1} \end{bmatrix}\right), K \equiv K_{RSGP}$$

2. Randomly sample t_p^n from the inferred mean and variance

3. Update the reward, reward-sensitive covariance K_{RS} , and hence K_{RSGP} based on t_p^n .

end

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 (ES, EL, HS, and HL) separately. To do so, we ordered t_p values associated with the same trial type within 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. For the same reason, the temporal distance in the kernel function was different from the actual trial lag (Figure 2B, S7). 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_{SE}^2 , and then used those fits to search for the best fit of the remaining hyperparameters (l_{RS} , σ_{RS} , and σ_0).

866 using all trials. The optimization of the multivariate likelihood function was achieved by line searching with
867 quadratic and cubic polynomial approximations. The conjugate gradients was used to compute the search
868 directions (Rasmussen and Williams, 2006). The landscape of likelihood indicated that the optimization was
869 convex for a wide range of initial values (Figure S6A, Table S1).

870 The RSGP model fit to data provides a prediction of the distribution of t_p on each trial based on previous
871 trials (t_p and reward history). We used this distribution to generate simulated values of t_p for each session,
872 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
873 same analysis we applied to behavioral data (Figure 5C). To derive an estimate of the slow part of error,
874 e_{slow} , we first fitted the RSGP to the behavior, and then used the reduced RSGP that only included the slow
875 kernel (K_{SE}) to predict the expected value of e_{slow} , i.e. the mean of a GP process governed by K_{SE} .

876 RSGP model for speed

877 We reformulated RSGP such that the key variable in the model was speed – not the interval (Figure S9 A).
878 Accordingly, the speed signal was used to control the slope of a ramping activity that triggered an action
879 when it breached a fixed threshold. The minor modification from target interval to speed enabled the RSGP
880 to capture both the dependence of variability on reward within each condition and the scalar variability across
881 conditions (Figure S9 B,C).

882 Markov chain Monte Carlo (MCMC) model

883 The MCMC model was proposed by Haith and Krakauer (Haith and Krakauer, 2014). We adapted this
884 algorithm to our task by adding time production noise to step 3, and approximating the cost as the opposite
885 of reward (i.e. $J(t_p) = -V(t_p)$). and simulated it to test whether it could capture (1) the long-term correlations,
886 (2) the monotonic relationship between the mean of error on trial n , $\mu(e^n)$, and error on trial $n-1$, e^{n-1} , and (3)
887 the U-shaped relationship between the standard deviation of error, $\sigma(e^n)$, and e^{n-1} . If we denote the variable
888 of interest, t_p , and the reward associated with it, $V(t_p)$, the algorithm they proposed works as follows:

-
1. Keep in memory a variable t_p^* that is currently associated with the highest value: $V(t_p^*)$
 2. On every trial, sample a new variable, denoted t_p^+ , in the vicinity of t_p^*
 3. Execute based on t_p^+ and measure $V(t_p^+)$
 4. Use a probabilistic Metropolis-Hastings rule to accept or reject t_p^+ as the new t_p^* depending on the relative values of $V(t_p^+)$, $V(t_p^*)$, and a free parameter β known in RL as the inverse temperature.

$$P(\text{Accept}) = \frac{e^{\beta \cdot V(t_p^+)}}{e^{\beta \cdot V(t_p^+)} + e^{\beta \cdot V(t_p^*)}}$$

5. Repeat step 2 - 4 till the end of the session.
-

889 **Table 4.** Algorithm for generating time series based on MCMC.

890

891 We have extensively explored the parameter space of this model. In the simulation (Figure S5 A), we used

0.1* t_t and 0.05* t_t as the level of sampling noise in step 2, and execution noise in step 3, respectively. The inverse temperature β was 100.

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 the previous section) using the following regression model:

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

where $\mathbf{r}$ represents a matrix ($n \times N$) containing spike counts within a 250 ms of N simultaneously recorded neurons across n trials, β_0 is a constant, and $\boldsymbol{\beta}$ is an 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 quantify the success of the model by computing the Pearson correlation coefficient (Figure 6B-D) 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).

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 $\boldsymbol{\beta}_{\text{Th}}$, and the other half to quantify the extent to which $\mathbf{z}_{\text{Th}}$ could predict e_{slow} . 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. Note that the number of sessions was combined across all 4 trial types as shown in Figure 6B-D bottom.

Statistical Analysis

Mean $\pm$ standard deviation (SD), Mean $\pm$ standard error of the mean (SEM) or median $\pm$ 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 (ratio of standard deviation to mean of t_p), (2) cross correlation between pairs of trials of different lag and trial type, (3) modulation of the variability by discrete or graded reward (4) variance terms in the RSGP model (σ_{SE}^2 , σ_{RS}^2 , and σ_0^2) 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 the variability of production interval for different pair-trial conditions (H0: equal variance).

Data Availability Statement and Accession Code Availability Statements

929 Data that support the findings of this study are available from the corresponding authors. Matlab codes for
 930 RSGP simulation and model fitting are available at <https://github.com/wangjing0/RSGP>.

931 **Mathematical notation**

932

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	The 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

933

934

935 **Author Contributions**

936 J.W., E.H. and M.J. conceived the project. J.W. and E.H. collected the main behavioral and
 937 electrophysiology data. J.W. analyzed the data and developed the model. E.H. played a major role in data
 938 analysis. J.W. and M.J. interpreted the data with contribution from E.H. and N.M. N.M. performed and
 939 analyzed the control experiment in monkeys. J.W. designed and analyzed the human psychophysics
 940 experiments and A.A. conducted the experiments. M.J. supervised the project. All authors contributed to
 941 the writing of the manuscript.

942

943 **Acknowledgments**

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

947

948 **Competing interests**

949 The authors declare no competing interests.

950

951 **Data and code availability**

952 The data and custom analysis scripts that support the findings of this study will be made available publicly
 953 on a designated website immediately after publication.

954 References

- 955 Afshar, A. *et al.* (2011) 'Single-trial neural correlates of arm movement preparation', *Neuron*, 71(3), pp. 555–564. doi:
956 10.1016/j.neuron.2011.05.047.
- 957 Ajemian, R. *et al.* (2013) 'A theory for how sensorimotor skills are learned and retained in noisy and nonstationary neural
958 circuits', *Proceedings of the National Academy of Sciences of the United States of America*, 110(52), pp. E5078–87. doi:
959 10.1073/pnas.1320116110.
- 960 Ames, K. C., Ryu, S. I. and Shenoy, K. V. (2014) 'Neural dynamics of reaching following incorrect or absent motor
961 preparation', *Neuron*, 81(2), pp. 438–451. doi: 10.1016/j.neuron.2013.11.003.
- 962 Ashmore, R. C. and Sommer, M. A. (2013) 'Delay activity of saccade-related neurons in the caudal dentate nucleus of the
963 macaque cerebellum', *Journal of neurophysiology*, 109(8), pp. 2129–2144. doi: 10.1152/jn.00906.2011.
- 964 van Beers, R. J. (2009) 'Motor learning is optimally tuned to the properties of motor noise', *Neuron*, 63(3), pp. 406–417.
965 doi: 10.1016/j.neuron.2009.06.025.
- 966 Berman, R. A. and Wurtz, R. H. (2011) 'Signals conveyed in the pulvinar pathway from superior colliculus to cortical area
967 MT', *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 31(2), pp. 373–384. doi:
968 10.1523/JNEUROSCI.4738-10.2011.
- 969 Box, G. E. P. *et al.* (2015) *Time Series Analysis: Forecasting and Control*. John Wiley & Sons. Available at:
970 <https://market.android.com/details?id=book-rNt5CgAAQBAJ>.
- 971 Carpenter, R. H. and Williams, M. L. (1995) 'Neural computation of log likelihood in control of saccadic eye movements',
972 *Nature*. Nature Publishing Group, 377(6544), pp. 59–62. doi: 10.1038/377059a0.
- 973 Cashaback, J. G. A. *et al.* (2017) 'Dissociating error-based and reinforcement-based loss functions during sensorimotor
974 learning', *PLoS computational biology*, 13(7), p. e1005623. doi: 10.1371/journal.pcbi.1005623.
- 975 Cashaback, J. G. A. *et al.* (2019) 'The gradient of the reinforcement landscape influences sensorimotor learning', *PLoS*
976 *computational biology*, 15(3), p. e1006839. doi: 10.1371/journal.pcbi.1006839.
- 977 Chaisanguanthum, K. S., Shen, H. H. and Sabes, P. N. (2014) 'Motor variability arises from a slow random walk in neural
978 state', *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 34(36), pp. 12071–12080. doi:
979 10.1523/JNEUROSCI.3001-13.2014.
- 980 Chen, X., Mohr, K. and Galea, J. M. (2017) 'Predicting explorative motor learning using decision-making and motor
981 noise', *PLoS computational biology*, 13(4), p. e1005503. doi: 10.1371/journal.pcbi.1005503.
- 982 Chen, Y., Ding, M. and Kelso, J. A. S. (1997) 'Long memory processes (1/f α type) in human coordination', *Physical*
983 *review letters*. APS, 79(22), p. 4501. Available at: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.79.4501>.
- 984 Churchland, M. M. *et al.* (2010) 'Stimulus onset quenches neural variability: a widespread cortical phenomenon', *Nature*
985 *neuroscience*. Nature Publishing Group, 13(3), pp. 369–378. doi: 10.1038/nn.2501.
- 986 Churchland, M. M. *et al.* (2012) 'Neural population dynamics during reaching', *Nature*. Nature Publishing Group,
987 487(7405), pp. 51–56. doi: 10.1038/nature11129.
- 988 Churchland, M. M., Afshar, A. and Shenoy, K. V. (2006) 'A central source of movement variability', *Neuron*, 52(6), pp.
989 1085–1096. doi: 10.1016/j.neuron.2006.10.034.
- 990 Church, R. M. and Broadbent, H. A. (1990) 'Alternative representations of time, number, and rate', *Cognition*, 37(1-2), pp.
991 55–81. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/2269008>.
- 992 Cohen, M. R. and Maunsell, J. H. R. (2009) 'Attention improves performance primarily by reducing interneuronal
993 correlations', *Nature neuroscience*, 12(12), pp. 1594–1600. doi: 10.1038/nn.2439.
- 994 Crossman, E. R. F. W. (1959) 'A THEORY OF THE ACQUISITION OF SPEED-SKILL*', *Ergonomics*. Taylor & Francis,
995 2(2), pp. 153–166. doi: 10.1080/00140135908930419.
- 996 Dam, G., Kording, K. and Wei, K. (2013) 'Credit Assignment during Movement Reinforcement Learning', *PloS one*. Public
997 Library of Science, 8(2), p. e55352. doi: 10.1371/journal.pone.0055352.

- 998 Daw, N. D. *et al.* (2006) 'Cortical substrates for exploratory decisions in humans', *Nature*, 441(7095), pp. 876–879. doi:
999 10.1038/nature04766.
- 1000 Daw, N. D., Niv, Y. and Dayan, P. (2005) 'Uncertainty-based competition between prefrontal and dorsolateral striatal
1001 systems for behavioral control', *Nature neuroscience*, 8(12), pp. 1704–1711. doi: 10.1038/nn1560.
- 1002 Dayan, P. and Daw, N. D. (2008) 'Decision theory, reinforcement learning, and the brain', *Cognitive, affective &*
1003 *behavioral neuroscience*, 8(4), pp. 429–453. doi: 10.3758/CABN.8.4.429.
- 1004 Dhawale, A. K. *et al.* (2019) 'Adaptive Regulation of Motor Variability', *Current biology: CB*. doi:
1005 10.1016/j.cub.2019.08.052.
- 1006 Dhawale, A. K., Smith, M. A. and Ölveczky, B. P. (2017) 'The Role of Variability in Motor Learning', *Annual review of*
1007 *neuroscience*. doi: 10.1146/annurev-neuro-072116-031548.
- 1008 Ebitz, R. B., Albarran, E. and Moore, T. (2018) 'Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in
1009 Prefrontal Cortex', *Neuron*, 97(2), p. 475. doi: 10.1016/j.neuron.2018.01.011.
- 1010 Faisal, A., Selen, L. and Wolpert, D. (2008) 'Noise in the nervous system', *Nature reviews. Neuroscience*, 9(4), pp. 292–
1011 303.
- 1012 Fee, M. S. and Goldberg, J. H. (2011) 'A hypothesis for basal ganglia-dependent reinforcement learning in the songbird',
1013 *Neuroscience*, 198, pp. 152–170. doi: 10.1016/j.neuroscience.2011.09.069.
- 1014 Frank, M. J. *et al.* (2009) 'Prefrontal and striatal dopaminergic genes predict individual differences in exploration and
1015 exploitation', *Nature neuroscience*, 12(8), pp. 1062–1068. doi: 10.1038/nn.2342.
- 1016 Gao, P. *et al.* (2017) 'A theory of multineuronal dimensionality, dynamics and measurement', *bioRxiv*. doi:
1017 10.1101/214262.
- 1018 Gibbon, J. (1977) 'Scalar expectancy theory and Weber's law in animal timing', *Psychological review. American*
1019 *Psychological Association*, 84(3), p. 279. doi: 10.1037/0033-295X.84.3.279.
- 1020 Gibbon, J., Church, R. M. and Meck, W. H. (1984) 'Scalar Timing in Memory', *Annals of the New York Academy of*
1021 *Sciences*. Blackwell Publishing Ltd, 423(1), pp. 52–77. doi: 10.1111/j.1749-6632.1984.tb23417.x.
- 1022 Gilden, D. L., Thornton, T. and Mallon, M. W. (1995) '1/f noise in human cognition', *Science*, 267(5205), pp. 1837–1839.
1023 doi: 10.1126/science.7892611.
- 1024 Gouvêa, T. S. *et al.* (2015) 'Striatal dynamics explain duration judgments', *eLife*, 4. doi: 10.7554/eLife.11386.
- 1025 Grossberg, S. and Schmajuk, N. A. (1989) 'Neural dynamics of adaptive timing and temporal discrimination during
1026 associative learning', *Neural networks: the official journal of the International Neural Network Society*, 2(2), pp. 79–102.
1027 doi: 10.1016/0893-6080(89)90026-9.
- 1028 Guo, Z. V. *et al.* (2017) 'Maintenance of persistent activity in a frontal thalamocortical loop', *Nature*, 545(7653), pp. 181–
1029 186. doi: 10.1038/nature22324.
- 1030 Haith, A. M. and Krakauer, J. W. (2014) 'Motor learning by sequential sampling of actions', in *Translational and*
1031 *computational motor control*. American Society of Neurorehabilitation Washington, DC. Available at: [http://blam-](http://blam-lab.org/wp-content/uploads/2016/12/Haith_TCMC_2014_mcmc.pdf)
1032 [lab.org/wp-content/uploads/2016/12/Haith_TCMC_2014_mcmc.pdf](http://blam-lab.org/wp-content/uploads/2016/12/Haith_TCMC_2014_mcmc.pdf).
- 1033 Halassa, M. M. *et al.* (2014) 'State-dependent architecture of thalamic reticular subnetworks', *Cell*, 158(4), pp. 808–821.
1034 doi: 10.1016/j.cell.2014.06.025.
- 1035 Harris, C. M. and Wolpert, D. M. (1998) 'Signal-dependent noise determines motor planning', *Nature*. Nature Publishing
1036 Group, 394(6695), pp. 780–784. doi: 10.1038/29528.
- 1037 Harris, K. D. and Thiele, A. (2011) 'Cortical state and attention', *Nature reviews. Neuroscience*, 12(9), pp. 509–523. doi:
1038 10.1038/nn3084.
- 1039 Hauser, C. K. *et al.* (2018) 'Motor selection dynamics in FEF explain the reaction time variance of saccades to single
1040 targets', *eLife*, 7. doi: 10.7554/eLife.33456.
- 1041 Hayden, B. Y., Pearson, J. M. and Platt, M. L. (2011) 'Neuronal basis of sequential foraging decisions in a patchy

- 1042 environment', *Nature neuroscience*, 14(7), pp. 933–939. doi: 10.1038/nn.2856.
- 1043 Herzfeld, D. J. *et al.* (2015) 'Encoding of action by the Purkinje cells of the cerebellum', *Nature*. Nature Research,
1044 526(7573), pp. 439–442. doi: 10.1038/nature15693.
- 1045 Hoshi, E. *et al.* (2005) 'The cerebellum communicates with the basal ganglia', *Nature neuroscience*. Nature Publishing
1046 Group, 8(11), pp. 1491–1493. doi: 10.1038/nn1544.
- 1047 Huang, C. *et al.* (2019) 'Circuit Models of Low-Dimensional Shared Variability in Cortical Networks', *Neuron*, 101(2), pp.
1048 337–348.e4. doi: 10.1016/j.neuron.2018.11.034.
- 1049 Huang, V. S. *et al.* (2011) 'Rethinking motor learning and savings in adaptation paradigms: model-free memory for
1050 successful actions combines with internal models', *Neuron*, 70(4), pp. 787–801. doi: 10.1016/j.neuron.2011.04.012.
- 1051 Hubel, D. H. and Wiesel, T. N. (1962) 'Receptive fields, binocular interaction and functional architecture in the cat's visual
1052 cortex', *The Journal of physiology*, 160, pp. 106–154. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/14449617>.
- 1053 Huberdeau, D. M., Krakauer, J. W. and Haith, A. M. (2015) 'Dual-process decomposition in human sensorimotor
1054 adaptation', *Current opinion in neurobiology*, 33, pp. 71–77. doi: 10.1016/j.conb.2015.03.003.
- 1055 Ito, M. (2002) 'Historical review of the significance of the cerebellum and the role of Purkinje cells in motor learning',
1056 *Annals of the New York Academy of Sciences*, 978, pp. 273–288. Available at:
1057 <https://www.ncbi.nlm.nih.gov/pubmed/12582060>.
- 1058 Izawa, J. and Shadmehr, R. (2011) 'Learning from sensory and reward prediction errors during motor adaptation', *PLoS*
1059 *computational biology*, 7(3), p. e1002012. doi: 10.1371/journal.pcbi.1002012.
- 1060 Jazayeri, M. and Shadlen, M. N. (2010) 'Temporal context calibrates interval timing', *Nature neuroscience*. Nature
1061 Publishing Group, 13(8), pp. 1020–1026. doi: 10.1038/nn.2590.
- 1062 Jazayeri, M. and Shadlen, M. N. (2015) 'A Neural Mechanism for Sensing and Reproducing a Time Interval', *Current*
1063 *biology: CB*. Elsevier, 25(20), pp. 2599–2609. doi: 10.1016/j.cub.2015.08.038.
- 1064 Joiner, W. M. and Smith, M. A. (2008) 'Long-term retention explained by a model of short-term learning in the adaptive
1065 control of reaching', *Journal of neurophysiology*, 100(5), pp. 2948–2955. doi: 10.1152/jn.90706.2008.
- 1066 Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996) 'Reinforcement Learning: A Survey', *Journal of Artificial*
1067 *Intelligence Research*, 4, pp. 237–285. doi: 10.1613/jair.301.
- 1068 Kao, M. H., Doupe, A. J. and Brainard, M. S. (2005) 'Contributions of an avian basal ganglia–forebrain circuit to real-time
1069 modulation of song', *Nature*. Macmillan Magazines Ltd., 433, p. 638. doi: 10.1038/nature03127.
- 1070 Karlsson, M. P., Tervo, D. G. R. and Karpova, A. Y. (2012) 'Network resets in medial prefrontal cortex mark the onset of
1071 behavioral uncertainty', *Science*, 338(6103), pp. 135–139. doi: 10.1126/science.1226518.
- 1072 Kato, H. K. *et al.* (2012) 'Dynamic sensory representations in the olfactory bulb: modulation by wakefulness and
1073 experience', *Neuron*, 76(5), pp. 962–975. doi: 10.1016/j.neuron.2012.09.037.
- 1074 Killeen, P. R. and Fetterman, J. G. (1988) 'A behavioral theory of timing', *Psychological review*, 95(2), pp. 274–295.
1075 Available at:
1076 <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=3375401&retmode=ref&cmd=prlinks>.
- 1077 van der Kooij, K. and Smeets, J. B. J. (2019) 'Reward-based motor adaptation can generalize across actions', *Journal of*
1078 *experimental psychology. Learning, memory, and cognition*, 45(1), pp. 71–81. doi: 10.1037/xlm0000573.
- 1079 Kunitatsu, J. *et al.* (2018) 'Different contributions of preparatory activity in the basal ganglia and cerebellum for self-
1080 timing', *eLife*, 7. doi: 10.7554/eLife.35676.
- 1081 Kunitatsu, J. and Tanaka, M. (2016) 'Striatal dopamine modulates timing of self-initiated saccades', *Neuroscience*. doi:
1082 10.1016/j.neuroscience.2016.09.006.
- 1083 Laming, D. (1979) 'Autocorrelation of choice-reaction times', *Acta psychologica*, 43(5), pp. 381–412. doi: 10.1016/0001-
1084 6918(79)90032-5.
- 1085 Lara, A. H. *et al.* (2018) 'Conservation of preparatory neural events in monkey motor cortex regardless of how movement

- 1086 is initiated', *eLife*, 7. doi: 10.7554/eLife.31826.
- 1087 Lau, B. and Glimcher, P. W. (2007) 'Action and outcome encoding in the primate caudate nucleus', *The Journal of*
1088 *neuroscience: the official journal of the Society for Neuroscience*, 27(52), pp. 14502–14514. doi:
1089 10.1523/JNEUROSCI.3060-07.2007.
- 1090 Lauwereyns, J. *et al.* (2002) 'A neural correlate of response bias in monkey caudate nucleus', *Nature*. Nature Publishing
1091 Group, 418(6896), pp. 413–417. doi: 10.1038/nature00892.
- 1092 Lee, M. D. *et al.* (2011) 'Psychological models of human and optimal performance in bandit problems', *Cognitive systems*
1093 *research*, 12(2), pp. 164–174. doi: 10.1016/j.cogsys.2010.07.007.
- 1094 Lee, S.-H. and Dan, Y. (2012) 'Neuromodulation of brain states', *Neuron*, 76(1), pp. 209–222. doi:
1095 10.1016/j.neuron.2012.09.012.
- 1096 Luck, S. J. *et al.* (1997) 'Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual
1097 cortex', *Journal of neurophysiology*, 77(1), pp. 24–42. doi: 10.1152/jn.1997.77.1.24.
- 1098 Machado, A. (1997) 'Learning the temporal dynamics of behavior', *Psychological review*, 104(2), pp. 241–265. Available
1099 at: <https://www.ncbi.nlm.nih.gov/pubmed/9127582>.
- 1100 Malapani, C. and Fairhurst, S. (2002) 'Scalar Timing in Animals and Humans', *Learning and motivation*. Elsevier, 33(1),
1101 pp. 156–176. doi: 10.1006/lmot.2001.1105.
- 1102 Massi, B., Donahue, C. H. and Lee, D. (2018) 'Volatility Facilitates Value Updating in the Prefrontal Cortex', *Neuron*,
1103 99(3), pp. 598–608.e4. doi: 10.1016/j.neuron.2018.06.033.
- 1104 Mauk, M. D. and Buonomano, D. V. (2004) 'The neural basis of temporal processing', *Annual review of neuroscience*, 27,
1105 pp. 307–340. doi: 10.1146/annurev.neuro.27.070203.144247.
- 1106 Mcalonan, K., Cavanaugh, J. and Wurtz, R. H. (2008) 'Guarding the gateway to cortex with attention in visual thalamus',
1107 *Nature*. Nature Publishing Group, 456(7220), pp. 391–394. doi: 10.1038/nature07382.
- 1108 Medina, J. and Lisberger, S. (2008) 'Links from complex spikes to local plasticity and motor learning in the cerebellum of
1109 awake-behaving monkeys', *Nature neuroscience*. Nature Publishing Group, 11(10), pp. 1185–1192. Available at:
1110 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18806784.
- 1111 Merchant, H. and Averbeck, B. B. (2017) 'The Computational and Neural Basis of Rhythmic Timing in Medial Premotor
1112 Cortex', *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 37(17), pp. 4552–4564. doi:
1113 10.1523/JNEUROSCI.0367-17.2017.
- 1114 Merrill, W. J., Jr and Bennett, C. A. (1956) 'The application of temporal correlation techniques in psychology', *The Journal*
1115 *of applied psychology*. American Psychological Association, 40(4), p. 272. Available at:
1116 <http://psycnet.apa.org/record/1957-05355-001>.
- 1117 Middleton, F. A. and Strick, P. L. (2000) 'Basal ganglia and cerebellar loops: motor and cognitive circuits', *Brain and*
1118 *cognition*, 42(2-3), pp. 183–200. doi: 10.1006/brcg.1999.1099.
- 1119 Mitchell, J. F., Sundberg, K. A. and Reynolds, J. H. (2009) 'Spatial attention decorrelates intrinsic activity fluctuations in
1120 macaque area V4', *Neuron*, 63(6), pp. 879–888. doi: 10.1016/j.neuron.2009.09.013.
- 1121 Murakami, M. *et al.* (2014) 'Neural antecedents of self-initiated actions in secondary motor cortex', *Nature neuroscience*,
1122 17(11), pp. 1574–1582. doi: 10.1038/nn.3826.
- 1123 Murakami, M. *et al.* (2017) 'Distinct Sources of Deterministic and Stochastic Components of Action Timing Decisions in
1124 Rodent Frontal Cortex', *Neuron*, 94(4), pp. 908–919.e7. doi: 10.1016/j.neuron.2017.04.040.
- 1125 Narain, D. *et al.* (2018) 'A cerebellar mechanism for learning prior distributions of time intervals', *Nature communications*,
1126 9(1), p. 469. doi: 10.1038/s41467-017-02516-x.
- 1127 Narayanan, N. S. and Laubach, M. (2008) 'Neuronal correlates of post-error slowing in the rat dorsomedial prefrontal
1128 cortex', *Journal of neurophysiology*, 100(1), pp. 520–525. doi: 10.1152/jn.00035.2008.
- 1129 Ni, A. M. *et al.* (2018) 'Learning and attention reveal a general relationship between population activity and behavior',
1130 *Science*, 359(6374), pp. 463–465. doi: 10.1126/science.aao0284.

- 1131 Niell, C. M. and Stryker, M. P. (2010) 'Modulation of visual responses by behavioral state in mouse visual cortex', *Neuron*,
1132 65(4), pp. 472–479. doi: 10.1016/j.neuron.2010.01.033.
- 1133 Nikooyan, A. A. and Ahmed, A. A. (2015) 'Reward feedback accelerates motor learning', *Journal of neurophysiology*,
1134 113(2), pp. 633–646. doi: 10.1152/jn.00032.2014.
- 1135 Olveczky, B. P., Andalman, A. S. and Fee, M. S. (2005) 'Vocal experimentation in the juvenile songbird requires a basal
1136 ganglia circuit', *PLoS biology*, 3(5), p. e153. doi: 10.1371/journal.pbio.0030153.
- 1137 Oprisan, S. A. and Buhusi, C. V. (2014) 'What is all the noise about in interval timing?', *Philosophical transactions of the
1138 Royal Society of London. Series B, Biological sciences*, 369(1637), p. 20120459. doi: 10.1098/rstb.2012.0459.
- 1139 Palidis, D. J., Cashaback, J. G. A. and Gribble, P. L. (no date) 'Neural Signatures of Reward and Sensory Prediction
1140 Error in Motor Learning'. doi: 10.1101/262576.
- 1141 Paton, J. J. and Buonomano, D. V. (2018) 'The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions',
1142 *Neuron*, 98(4), pp. 687–705. doi: 10.1016/j.neuron.2018.03.045.
- 1143 Pekny, S. E., Izawa, J. and Shadmehr, R. (2015) 'Reward-dependent modulation of movement variability', *The Journal of
1144 neuroscience: the official journal of the Society for Neuroscience*, 35(9), pp. 4015–4024. doi: 10.1523/JNEUROSCI.3244-
1145 14.2015.
- 1146 Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian process for machine learning*. MIT press.
- 1147 Remington, E. D., Egger, S. W., et al. (2018) 'A Dynamical Systems Perspective on Flexible Motor Timing', *Trends in
1148 cognitive sciences*, 22(10), pp. 938–952. doi: 10.1016/j.tics.2018.07.010.
- 1149 Remington, E. D., Narain, D., et al. (2018) 'Flexible Sensorimotor Computations through Rapid Reconfiguration of
1150 Cortical Dynamics', *Neuron*. Elsevier, 98(5), pp. 1005–1019.e5. doi: 10.1016/j.neuron.2018.05.020.
- 1151 Ruff, D. A. and Cohen, M. R. (2014) 'Attention can either increase or decrease spike count correlations in visual cortex',
1152 *Nature neuroscience*, 17(11), pp. 1591–1597. doi: 10.1038/nn.3835.
- 1153 Saalman, Y. B. et al. (2012) 'The pulvinar regulates information transmission between cortical areas based on attention
1154 demands', *Science*, 337(6095), pp. 753–756. doi: 10.1126/science.1223082.
- 1155 Santos, F. J. et al. (2015) 'Corticostriatal dynamics encode the refinement of specific behavioral variability during skill
1156 learning', *eLife*, 4, p. e09423. doi: 10.7554/eLife.09423.
- 1157 Sarafyazd, M. and Jazayeri, M. (2019) 'Hierarchical reasoning by neural circuits in the frontal cortex', *Science*, 364(6441).
1158 doi: 10.1126/science.aav8911.
- 1159 Schmitt, L. I. et al. (2017) 'Thalamic amplification of cortical connectivity sustains attentional control', *Nature*, 545(7653),
1160 pp. 219–223. doi: 10.1038/nature22073.
- 1161 Sheahan, H. R., Franklin, D. W. and Wolpert, D. M. (2016) 'Motor Planning, Not Execution, Separates Motor Memories',
1162 *Neuron*, 92(4), pp. 773–779. doi: 10.1016/j.neuron.2016.10.017.
- 1163 Shmuelof, L. et al. (2012) 'Overcoming Motor "Forgetting" Through Reinforcement Of Learned Actions', *The Journal of
1164 neuroscience: the official journal of the Society for Neuroscience*. Society for Neuroscience, 32(42), pp. 14617–14621a.
1165 doi: 10.1523/JNEUROSCI.2184-12.2012.
- 1166 Shmuelof, L., Krakauer, J. W. and Mazzoni, P. (2012) 'How is a motor skill learned? Change and invariance at the levels
1167 of task success and trajectory control', *Journal of neurophysiology*, 108(2), pp. 578–594. doi: 10.1152/jn.00856.2011.
- 1168 Simen, P. et al. (2011) 'A model of interval timing by neural integration', *The Journal of neuroscience: the official journal
1169 of the Society for Neuroscience*, 31(25), pp. 9238–9253. doi: 10.1523/JNEUROSCI.3121-10.2011.
- 1170 Smith, M. A., Ghazizadeh, A. and Shadmehr, R. (2006) 'Interacting adaptive processes with different timescales underlie
1171 short-term motor learning', *PLoS biology*, 4(6), p. e179. doi: 10.1371/journal.pbio.0040179.
- 1172 Sommer, M. A. and Wurtz, R. H. (2006) 'Influence of the thalamus on spatial visual processing in frontal cortex', *Nature*.
1173 Nature Publishing Group, 444(7117), pp. 374–377. doi: 10.1038/nature05279.
- 1174 Staddon, J. E. and Higa, J. J. (1999) 'Time and memory: towards a pacemaker-free theory of interval timing', *Journal of*

- 1175 *the experimental analysis of behavior*, 71(2), pp. 215–251. doi: 10.1901/jeab.1999.71-215.
- 1176 Sternad, D. and Abe, M. O. (2010) 'Variability, Noise, and Sensitivity to Error in Learning a Motor Task', in *Motor Control*.
1177 Oxford University Press. doi: 10.1093/acprof:oso/9780195395273.003.0012.
- 1178 Sul, J. H. *et al.* (2010) 'Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making', *Neuron*,
1179 66(3), pp. 449–460. doi: 10.1016/j.neuron.2010.03.033.
- 1180 Sutton, R. S. and Barto, A. G. (1998) *Reinforcement Learning: An Introduction*. MIT Press. Available at:
1181 <https://market.android.com/details?id=book-CAFR6IBF4xYC>.
- 1182 Takikawa, Y., Kawagoe, R. and Hikosaka, O. (2002) 'Reward-dependent spatial selectivity of anticipatory activity in
1183 monkey caudate neurons', *Journal of neurophysiology*. American Physiological Society, 87(1), pp. 508–515.
- 1184 Tervo, D. G. R. *et al.* (2014) 'Behavioral variability through stochastic choice and its gating by anterior cingulate cortex',
1185 *Cell*, 159(1), pp. 21–32. doi: 10.1016/j.cell.2014.08.037.
- 1186 Thoroughman, K. A. and Shadmehr, R. (2000) 'Learning of action through adaptive combination of motor primitives',
1187 *Nature*. Nature Publishing Group, 407(6805), pp. 742–747. doi: 10.1038/35037588.
- 1188 Tumer, E. C. and Brainard, M. S. (2007) 'Performance variability enables adaptive plasticity of "crystallized" adult
1189 birdsong', *Nature*, 450(7173), pp. 1240–1244. doi: 10.1038/nature06390.
- 1190 Vaswani, P. A. *et al.* (2015) 'Persistent residual errors in motor adaptation tasks: reversion to baseline and exploratory
1191 escape', *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 35(17), pp. 6969–6977. doi:
1192 10.1523/JNEUROSCI.2656-14.2015.
- 1193 Verstynen, T. and Sabes, P. N. (2011) 'How each movement changes the next: an experimental and theoretical study of
1194 fast adaptive priors in reaching', *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 31(27),
1195 pp. 10050–10059. doi: 10.1523/JNEUROSCI.6525-10.2011.
- 1196 Vinck, M. *et al.* (2015) 'Arousal and locomotion make distinct contributions to cortical activity patterns and visual
1197 encoding', *Neuron*, 86(3), pp. 740–754. doi: 10.1016/j.neuron.2015.03.028.
- 1198 Vyas, S. *et al.* (2018) 'Neural Population Dynamics Underlying Motor Learning Transfer', *Neuron*, 97(5), pp. 1177–
1199 1186.e3. doi: 10.1016/j.neuron.2018.01.040.
- 1200 Wagenmakers, E.-J., Farrell, S. and Ratcliff, R. (2004) 'Estimation and interpretation of 1/ α noise in human
1201 cognition', *Psychonomic bulletin & review*, 11(4), pp. 579–615. Available at:
1202 <http://www.ncbi.nlm.nih.gov/pubmed/15581115>.
- 1203 Wang, J. *et al.* (2018) 'Flexible timing by temporal scaling of cortical responses', *Nature neuroscience*, 21(1), pp. 102–
1204 110. doi: 10.1038/s41593-017-0028-6.
- 1205 Weiss, B., Coleman, P. D. and Green, R. F. (1955) 'A stochastic model for time-ordered dependencies in continuous
1206 scale repetitive judgments', *Journal of experimental psychology*. American Psychological Association, 50(4), p. 237.
1207 Available at: <http://psycnet.apa.org/record/1956-05580-001>.
- 1208 Wilson, R. C. *et al.* (2014) 'Humans use directed and random exploration to solve the explore-exploit dilemma', *Journal of*
1209 *experimental psychology. General*, 143(6), pp. 2074–2081. doi: 10.1037/a0038199.
- 1210 Wimmer, R. D. *et al.* (2015) 'Thalamic control of sensory selection in divided attention', *Nature*, 526(7575), pp. 705–709.
1211 doi: 10.1038/nature15398.
- 1212 Wolpert, D. M., Diedrichsen, J. and Flanagan, J. R. (2011) 'Principles of sensorimotor learning', *Nature reviews*.
1213 *Neuroscience*, 12(12), pp. 739–751. doi: 10.1038/nrn3112.
- 1214 Wu, H. G. *et al.* (2014) 'Temporal structure of motor variability is dynamically regulated and predicts motor learning
1215 ability', *Nature neuroscience*, 17(2), pp. 312–321. doi: 10.1038/nn.3616.
- 1216 Xiong, Q., Znamenskiy, P. and Zador, A. M. (2015) 'Selective corticostriatal plasticity during acquisition of an auditory
1217 discrimination task', *Nature*. doi: 10.1038/nature14225.
- 1218 Yasuda, M. and Hikosaka, O. (2015) 'Functional territories in primate substantia nigra pars reticulata separately signaling
1219 stable and flexible values', *Journal of neurophysiology*, 113(6), pp. 1681–1696. doi: 10.1152/jn.00674.2014.

1220 Zhou, H., Schafer, R. J. and Desimone, R. (2016) 'Pulvinar-Cortex Interactions in Vision and Attention', *Neuron*, 89(1),
 1221 pp. 209–220. doi: 10.1016/j.neuron.2015.11.034.

1222

1223 **Supplementary information**

1224 **Reinforcement regulates timing variability in thalamus**

1225 Jing Wang^{1,2}, Eghbal Hosseini², Nicolas Meirhaeghe³, Adam Akkad², Mehrdad Jazayeri²

1226 ¹Department of Bioengineering, University of Missouri, Columbia, MO 65211 ²McGovern Institute for Brain
 1227 Research, Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge,
 1228 Massachusetts 02139 ³Harvard-MIT Division of Health Sciences & Technology, Cambridge, Massachusetts
 1229 02139, USA

1230

1231 **Figure S1. Dependence of timing variability on reward rate.** The gray scale shows the marginal distribution of $\sigma(t_p)$ for given $\mu(t_p)$
 1232 derived from local estimates of $\sigma(t_p)$ and $\mu(t_p)$ using a 50-trial sliding window (yellow dots: mean of $\sigma(t_p)$ for each $\mu(t_p)$ value; colored
 1233 dash lines: target intervals centered at 800 and 1500 ms). Bin size was 20 ms.

1234

1235

1236 **Figure S2. Context-dependent slow fluctuations of timing variability for (A) transitions between different trial types and for**
 1237 **(B) different saccade directions.** Same as Figure 2B shown for the two animals and different trial types separately. (B) Similar to
 1238 Figure 2B, it shows the long range correlation between trials with same and different saccade direction separately. The analysis was
 1239 performed separately for trials of the same type (i.e., same target interval and same effector), and pooled afterward (black: same
 1240 direction; purple: different direction; bold: average; dashed [1%-99%] confidence intervals). The long-term correlations were not
 1241 significantly different for the same and different directions (Monkeys: paired sample t-test, $p = 0.39$, $t = 0.86$, $df = 159$. Humans:
 1242 paired sample t-test, $p = 0.55$, $t = 0.60$, $df = 79$).

Figure S3. Serial correlation of t_p in the control experiment. Slow fluctuations of t_p were diminished in the control experiment in which the demand for remembering the target interval was minimized. Results are shown in the same format as in Figure 2A.

1248

1249 **Figure S4. The effect of reward on behavioral variability.** (A) Cross-condition relationship between produced interval (t_p) in
1250 consecutive trials. The results for each animal are organized on a 4x4 panel covering the full set of 16 possible transitions between
1251 4 trial types (ES: Eye-Short; EL: Eye-Long; HS: Hand-Short; HL: Hand-Long) with the abscissa and ordinate showing t_p at trial $n-1$
1252 (t_p^{n-1}) and n (t_p^n), respectively. In each panel, the median t_p^n as the function of t_p^{n-1} is marked in bold line at the center. The 25th and
1253 75th percentile are marked with medium thickness, and the 10th and 90th percentile with thin lines. Bin size was 20 ms. (B) Error
1254 statistics within each session. Since our original analysis was based on data combined across sessions, one potential concern we
1255 had was that modulation of t_p variance with reward was due to differences across (but not within) sessions. To address this, we first
1256 removed outlier t_p values (more than 3 s.d. away from the mean) and then analyzed z-scored errors within each session (denoted $\tilde{e}$
1257). This normalization ensured that error pairs in every session were drawn from a zero-mean and unit-variance distribution. Results
1258 are shown in the same format as in Figure 3B and C. The variability increased significantly after unrewarded trials compared to that
1259 after rewarded trials (horizontal lines, *** $p < 0.001$; monkey A: $F(8889,10621) = 1.15$ for negative $\tilde{e}$ and $F(9841,11103) = 1.09$ for
1260 positive $\tilde{e}$; monkey D: $F(21216,12364) = 1.25$ for negative $\tilde{e}$ and $F(1600,19293) = 1.38$ for positive $\tilde{e}$), and this effect was context
1261 specific (vertical lines, for monkey A: $F(10621,9841) = 1.10$ and monkey D: $F(12364, 16100) = 1.11$). (C) Behavioral variability as a
1262 function of the magnitude of reward in the preceding trial. The distribution of $\sigma(e^n)$ sorted into three bins depending on reward
1263 magnitude (black: no reward; dark green: low reward; light green: high reward). The variability was significantly larger in the absence
1264 of reward in comparison to low reward (***, $p < .001$, two-sample t-test, $t = 6.6$, $df = 117$), and significantly larger in the low compared
1265 to the high reward (**, $p = 0.008$, two-sample t-test, $t = 2.5$, $df = 68$). Results were generated using bootstrapping; i.e., we computed
1266 $\sigma(e^n)$ from randomly sampled trials without replacement within each bin of e^{n-1} .

1267

1268

Figure S5. Analysis of error statistics for alternative generative processes and the causal experiment. (A) A Markovian sampling model. Production interval (t_p) was generated as samples from a Markov chain Monte Carlo (MCMC) model (see Methods). **(B)** This model does not exhibit the long-term autocorrelations observed in animals' behavior. Moreover, when the model was parameterized to capture the monotonic relationship between $\mu(e^n)$ and e^{n-1} , it failed to capture the U-shaped relationship between $\sigma(e^n)$ and e^{n-1} ($s_2 = 0.045$ [-0.048 0.14]). Example data from Monkey D is shown in orange. **(C)** The effect of reward on behavioral variability in the probabilistic reward experiment. Result was combined for all subjects in Fig. 4D and was plotted in the same format as Fig. 3B. To avoid sampling bias, same number of trials were drawn repeatedly after "incorrect" (red, -Rew $^{n-1}$) or "correct" (green, +Rew $^{n-1}$) trials (Shaded area: standard error of the mean computed from 100 bootstraps). Behavioral variability $\sigma(e^n)$ increased significantly after 'incorrect' trial outcome regardless of the size of error (***) $p < 0.001$, two-sample t-test on $\sigma(e^n)$ between the two conditions, $t_{198} > 6.70$ for all bins). The slope of the $\mu(e^n)$ as a function of e^{n-1} was larger after 'correct' outcome (***) $p < 0.001$, two sample t-test $t_{198} = 16.8$) demonstrating stronger behavioral correlation after positive reinforcement. **(D)** Same as (C) for the behavior of the RSGP model fit to the subject's behavior.

1283

1284

1285 **Figure S6. Long-term correlation of t_p and context-specificity in humans.** Correlation was averaged across human subjects (N
1286 = 5). Same format as in Figure 2A and B.

1287

1288

1289 **Figure S7. Error statistics as a function of trial lag. (A)** Top: Average error on trial n , $\mu(e^n)$ as a function of error in trial
1290 $n-k$ (e^{n-k}) for $k = 0, 1, 2$, and 20 (different colors). Bottom: Standard deviation of error on trial n , $\sigma(e^n)$ as a function of e^{n-k} .
1291 Results were computed for each trial type separately averaged afterwards. **(B)** Modulation of variability by reward as a
1292 function of trial lag. Similar to the main paper, we tested the U-shaped profile using quadratic regression. The ordinate
1293 shows the coefficient of the square term (s_2) of the quadratic regression as a function of trial lag (circles: s_2 for the k values
1294 in A; shaded area: [1%-99%] confidence intervals). The reward has no direct bearing on the variability of the current trial
1295 ($k=0$), and its effect drops as a function of trial lag.

Figure S8. Results of the RSGP model fit to the behavior. (A) Simulation results indicated that the likelihood function captures the ground truth and that the error surface is convex for a wide range of initial values. See Methods for details. The hyperparameters used for simulation are marked by asterisks and the results of fitting are in Table S1. (B) Same as Figure 5B shown for the two animals separately. (C) The model fit of variances (σ_{SE} , σ_{RS} and σ_0) across sessions for monkeys (left column) and humans (right column).

	l_{SE}	σ_{SE}	l_{RS}	σ_{RS}	σ_0
Ground truth	20.0	0.141	2.0	0.141	0.10
MML fit	18.4	0.129	2.14	0.137	0.0749

Table S1. Ground truth versus model fits for the hyperparameters used in RSGP simulation in Figure 5A and S8A.

1306
1307

Figure S9. RSGP model for speed. (A) The model has two modules. The first module provides a tonic speed input, and the second module integrates the speed and generates a ramp toward a fixed (i.e. interval-independent) action-triggering threshold. The speed controls the slope of the ramp and therefore, the resulting produced interval (t_p). Similar to the task, the presence or absence of reward depends on the error in t_p relative to the desired interval. The RSGP controls the output of the first module, and governs both the slow fluctuations of the speed (i.e., drift across trials), and the sensitivity of the speed to previous trials' reward (reinforcement). Parameters used for generating speed were identical to those used previously in the RSGP (Figure 5A, Table S1). The only difference is that the model incorporates the instruction about the target interval, t_i , by applying a gain factor to the output of the first module that is inversely proportional to t_i . This ensures that, on average, the ramp would reach the threshold at the instructed time. (B) t_p distribution of the model in an example session for the Short and Long conditions. (C) The relationship between the mean and standard deviation of t_p based on simulations of the model. The results were produced using the same analysis used in Figure 1D. The model captures the increase of variability for the Long relative to Short condition (i.e., scalar variability) as evident from the slope of the regression line (solid line; positive slope, $p < 0.01$). It also captures the U-shaped sensitivity of variability to reward around the desired interval (dashed line) for each condition. (D) Partial correlation of t_p shows slow fluctuations.

1322

1323 **Figure S10. The relationship between neural activity on trial n to relative error in the preceding trial (e^{n-1}) in three brain**
1324 **areas and two animals. (A) and (B) Same format as in Figure 7B-D, with the addition of confidence intervals [1%-99%] (shaded**
1325 **region) obtained by randomly sampling a subset of trials in each bin of e^{n-1} .**

1326

Figure S11. Drift and reward-dependent variability in the three brain areas inferred from comparable number of simultaneously recorded neurons. (A) Histogram of the number of simultaneously recorded neurons in the thalamus, DMFC and caudate across sessions. The averages are shown on top. (B) The number of simultaneously recorded neurons in the thalamus subsampled by a factor of 2 to approximately match DMFC. (C) Analysis of drift and reward-dependent variability for the sub-sampled thalamic population, shown in the same format as Figure 7B.

1334

1335 **Figure S12. Drift and variability of spike count along the direction that decodes the produced interval (t_p) in the**
1336 **thalamus, DMFC and caudate. (A)** Histogram of the correlation coefficients between t_p and the output of a decoder of t_p
1337 from neural data, denoted z (black: significantly positive). Results are shown with the same format as in Figure 6B. **(B)**
1338 Mean (top; $\mu(z^n)$) and standard deviation (bottom; $\sigma(z^n)$) of projected neural activity as a function of error in the preceding
1339 trial (e^{n-1}). Results are shown with the same format as in Figure 7B (thick lines: same effector; thin lines: different effectors).
1340

Brain area	Parameters		Same effector	Different effector
Thalamus	$\mu(z^n)$	m_1	0.13 [0.096 0.17]	0.018 [0.002 0.036]
		m_0	0.012 [0.005 0.019]	0.006 [0.002 0.01]
	$\sigma(z^n)$	s_2	0.32 [0.19 0.44]	-0.019 [-0.07 0.066]
DMFC	$\mu(z^n)$	s_1	0.012 [-0.019 0.046]	-0.003 [-0.02 0.014]
		s_0	0.063 [0.055 0.071]	0.065 [0.061 0.069]
	$\sigma(z^n)$	s_2	0.15 [0.087 0.21]	-0.017 [-0.071 0.036]
Caudate	$\mu(z^n)$	s_1	-0.006 [-0.022 0.009]	0.001 [-0.0041 0.024]
		s_0	0.049 [0.045 0.052]	0.051 [0.047 0.054]
	$\sigma(z^n)$	s_2	0.077 [0.036 0.12]	0.007 [-0.012 0.027]
	$\mu(z^n)$	m_1	0.010 [0.0014 0.018]	0.015 [0.011 0.019]
		m_0		
	$\sigma(z^n)$	s_2	0.16 [0.11 0.21]	0.044 [0.016 0.072]
	$\mu(z^n)$	s_1	0.005 [-0.007 0.17]	-0.0025 [-0.01 0.0047]
		s_0	0.050 [0.047 0.053]	0.051 [0.049 0.053]
	$\sigma(z^n)$	s_2		

1341

1342 **Table S2. Regression model fits relating spike count along the direction that predicts produced interval (t_p) on trial**
1343 **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
1344 e^{n-1} ; i.e., $\mu(z^n) = m_0 + m_1 e^{n-1}$. s_0 , s_1 and s_2 are parameters of the quadratic regression model relating the standard deviation
1345 of z^n ($\sigma(z^n)$) to e^{n-1} ; i.e., $\sigma(z^n) = s_0 + s_1 e^{n-1} + s_2 (e^{n-1})^2$. Fit parameters are shown separately for the thalamus, DMFC and caudate
1346 and further separated depending on whether trial $n-1$ and n were of the same or different effectors. Bold indicated
1347 significantly positive value ($p < 0.01$, 1% and 99% confidence intervals of the estimation were shown).

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

1566

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

1620

1621

1622

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

1674

1675

1676

1677

1678

1353

1354 **Figure S14. Analysis of neural data in relation to saccade direction. (A)** Error statistics across consecutive trials. $\mu(e^n)$
1355 (top) and $\sigma(e^n)$ (bottom) as a function of e^{n-1} . Similar to the main paper, we tested the U-shaped profile using quadratic
1356 regression. The coefficients of the squared term was not significantly different between the two directions (Monkeys: 0.26
1357 [0.17, 0.34] vs 0.27 [0.2, 0.34]; Humans: 0.33 [0.063, 0.59] vs 0.4 [0.24, 0.55]). **(B)** A Venn diagram showing the number
1358 of individual neurons in three brain areas whose spiking during a 250 ms time window before Set was significantly different
1359 across effectors (Eye/Hand; $p < 0.01$, two-sample t -test), target intervals (Short/Long; $p < 0.01$, two-sample t -test), saccade
1360 directions (Left/Right; $p < 0.01$, two-sample t -test), and predictive of the produced interval ($p < 0.01$, shuffling test). The
1361 number of neurons in each category is indicated in the diagram. Among the neurons that were predictive of the produced
1362 interval (intersection of blue and dark red), 3/320 in the thalamus, 17/191 in DMFC, and 1 out of 85 in the caudate were
1363 direction selective.

1364
1365

1366 **Figure S15. Euclidean distance between spike counts across the population associated with different trial types.** Each
1367 dendrogram summarizes the Euclidean distance of population spike count between different trial types as shown below the abscissa
1368 (E/H: Eye/Hand effector, S/L: Short/Long interval, l/r: left/right saccade direction). Spike counts were obtained from a 250 ms time
1369 window before Set in each area (columns, the total number of neurons used were indicated) and each animal (rows). We assessed
1370 the reliability of the organization of the dendrogram by applying the same analysis 100 times to 20% of trials randomly sampled from
1371 a specific trial type. The structure of the top nodes associated with effector and interval were more consistent across repeats than
1372 expected by chance ($p < 0.01$). The plot shows the results for 5 such random repeats. Distances were normalized by the number of
1373 neurons in each area and animal (i.e. divided by $\sqrt{n}$ to obtain the average distance per neuron in the state space).

1376

1377 **Figure S16. Analysis of speed variability across single trials as a function of reward.** (A) Schematics showing how we
 1378 estimated speed of population neural trajectories for single trial (v^n). We estimated the average speed for each trial by averaging the
 1379 Euclidean distance of population spike counts between nearby points ($dt = 125$ ms) along the trajectory. To minimize the differential
 1380 contribution of neurons with different average firing rates to speed estimates, spike counts for each neuron was normalized (z-score)
 1381 across trials and time bins. (B) Variability of speed estimates ($\sqrt{\text{Var}(v^n)}$) after rewarded trials ($+\text{Rew}^{n-1}$, ordinate) compared to
 1382 unrewarded trials ($-\text{Rew}^{n-1}$, abscissa) for consecutive trials of the same (top row) and different effector (bottom row) in the thalamus
 1383 (left), DMFC (middle) and caudate (right). In every session, each trial type contributed two points to each panel, one for when the
 1384 preceding error was positive and one for when it was negative. We treated these two error conditions independently so as to minimize
 1385 any potential dependence of our estimated $\sqrt{\text{Var}(v^n)}$ on average speed v^n that may change depending on the sign of error (Wang
 1386 *et al.*, 2018). The scatter plot in each panel shows the distribution of speed variability for trials that succeeded a rewarded trial
 1387 (ordinate) relative to the distribution of speed variability for trials that succeeded an unrewarded trial (abscissa). To test whether the
 1388 speed was significantly different between these two conditions, we used a two-tail paired sample t-test to determine whether and in
 1389 which direction the data deviated significantly from the unity line (diagonal). The distribution of the difference between the two
 1390 conditions is shown on the top right of each panel, and the mean for each distribution is indicated (triangle). P-values for each test
 1391 are indicated. (Thalamus: same effector, $t_{271} = 0.94$, different effector, $t_{271} = -0.88$, DMFC: same effector, $t_{313} = 2.69$, different effector,
 1392 $t_{313} = 1.03$, Caudate: same effector, $t_{161} = 2.67$, different effector, $t_{161} = 1.47$).

1393

1394 **References**

1395

1396 Wang, J. *et al.* (2018) 'Flexible timing by temporal scaling of cortical responses', *Nature neuroscience*, 21(1), pp. 102–
1397 110. doi: 10.1038/s41593-017-0028-6.