Midbrain dopamine firing activity codes reward expectation and motivation in a parametric working memory task

Stefania Sarno [a,b,c,d], Manuel Beirán [e,a], Gabriel Diaz-deLeon [f], Román Rossi-Pool [f], Ranulfo Romo [f,g] and Néstor Parga [a,b]

a: Departamento de Física Teórica, Universidad Autónoma de Madrid, Cantoblanco 28049, Madrid, Spain;

b: Centro de Investigación Avanzada en Física Fundamental, Universidad Autónoma de Madrid, Cantoblanco 28049, Madrid, Spain;

c: Aix Marseille Univ, INSERM, INMED. Turing Center for Living Systems, Marseille, France.

d: Aix Marseille Univ, CNRS, Centrale Marseille, IRPHE, Marseille, France.

e: Laboratoire de Neurosciences Cognitives Computationnelles, INSERM U960, École Normale Supérieure - PSL University, 75005 Paris, France;

f: Instituto de Fisiología Celular, Neurociencias, Universidad Nacional Autónoma de México, 04510 México DF, México;

g: El Colegio Nacional, 06020 México DF, México.

Abstract

Little is known about how dopamine (DA) neuron firing rates behave in cognitively demanding decision-making tasks. We investigated midbrain DA activity in monkeys performing a vibrotactile frequency discrimination task that required comparing two frequencies presented sequentially in time. We found that DA activity was involved in reward prediction, motivation and working memory (WM). Further, DA phasic responses to the stimuli were greatly affected by a contraction bias. They were also related to motivated behavior on a trial-by-trial basis, exhibiting a larger engagement in more difficult trials. Otherwise, dopamine WM activity was neither tuned to the initial stored frequency nor affected by the bias, although it did code the motivation level and exhibited a ramp-like increase. This result is consistent with a DA role in stabilizing sustained activity, by increasing the gain of prefrontal neurons in a motivation-dependent way. All these DA responses could be potentially related to cognitive control.

INTRODUCTION

Recent studies have shown that stimulation uncertainty affects the phasic responses of midbrain dopamine (DA) neurons to relevant cues during perceptual decision-making tasks. It was observed that those responses were influenced by cortical decision-making processes such as stimulus detection (de Lafuente and Romo, 2011) and evidence accumulation (Nomoto et al., 2010), in a manner consistent with the reward prediction error (RPE) hypothesis (Lak et al., 2017; Sarno et al., 2017). Importantly, phasic (fast fluctuations) DA responses exhibit beliefs about the state of the environment under uncertain conditions of stimulus perception (Babayan et al., 2018; Gershman and Uchida, 2019; Lak et al., 2017; Rao, 2010; Sarno et al., 2017; Starkweather et al., 2017), which are also modulated based on the confidence that the animal has in its own choice (de Lafuente and Romo, 2011; Lak et al., 2017; Sarno et al., 2017). These studies represent substantial progress towards understanding how DA neurons behave in the presence of uncertain stimulus conditions. However, little is known about phasic or tonic (fast or slow fluctuations) DA signaling involving the working memory (WM) delay-period or internal biases, specifically during a cognitively demanding task.

The vibrotactile frequency discrimination task facilitates deeper investigation into these current questions (Hernández et al., 1997; Romo et al., 1999). In this task, monkeys discriminate between two randomly selected vibrotactile stimulus frequencies, which were individually paired for each trial, and separated by 3 seconds (the WM delay, Fig. 1A). Importantly, the monkeys had to maintain a trace of the first stimulus (f_1) during the 3 second delay for comparison against the second stimulus (f_2) to reach a decision: f_1 is greater than f_2 ($f_1 > f_2$), or vice versa ($f_1 < f_2$). We sought to determine whether DA neurons were tuned to the value of f_1 or encoded reward-related information during the WM period. This is a question of interest, since DA neurons may

be associated to frontal area responses during WM. Specifically, frontal neurons were tuned to the identity of f_1 during the WM delay in a time-dependent manner (Brody et al., 2003; Romo and Rossi-Pool, 2020; Romo et al., 1999). The discrimination decision requires comparison of the two stimuli magnitudes, although it is likely that the perceptual evaluations are affected by internal processes (Körding, 2007; Preuschhof et al., 2010). In this case, the difficulty of the stimulus set (f_1, f_2 pairs) can be acutely different from the subjective difficulty perceived by the monkeys. Concordantly, monkeys shift perception of f_1 towards the center of the set's range (10-34 Hz) in our task, introducing a bias in the comparison between the two stimuli (Hollingworth, 1910). This is called contraction bias, and has been observed in many instances of perceptual discrimination in both humans and rodents, employing somatosensory or auditory tasks (Akrami et al., 2018; Ashourian and Loewenstein, 2011; Dyjas et al., 2012; Fassihi et al., 2014; Raviv et al., 2012).

Based on findings in DA neurons, the contraction bias could affect the activity of these neurons during stimulation periods and WM. In the vibrotactile discrimination task, the animal initiates an arm/hand movement after the offset of f_2 to report its decision by pushing one of two buttons; the time lag between f_2 offset and movement initiation is called the reaction time (RT, Fig. 1A). If we consider RTs as a way to measure motivation (Satoh et al., 2003) it can be used to study the temporal evolution of the DA activity. Some prior claims associate tonic DA activity with conveying motivation (Hamid et al., 2016; Niv et al., 2007), while others also correlated motivation with phasic activity (Berke, 2018; Satoh et al., 2003). We wonder whether motivation during the discrimination task relates to the transient responses to the stimuli, and whether motivation influences the DA working memory activity.

Our results found that the phasic DA responses were controlled by the contraction bias and not tuned to the frequency values, in contrast to the monotonic tuning reported previously in frontal areas (Brody et al., 2003; Romo et al., 1999). This bias induced a subjective difficulty that modulated the responses, decision reports, and RTs. A Bayesian model (Ashourian and Loewenstein, 2011; Preuschhof et al., 2010) gave a measure of the animal's decision confidence, which also appeared modulated by the bias. The phasic responses were negatively correlated with RTs on a trial-by-trial basis when the subjective task difficulty was held constant. This finding suggests that this DA signal was driving the motivation behind the monkeys' behavior. Interestingly, the DA signal also exhibited a ramping-like increase during the WM delay period. Although it is hard to assign a functional role to this DA signal, we propose that it could be involved in the internal control process of the animal's performance.

RESULTS

Contraction bias shapes behavioral performance

Monkeys were trained to discriminate between two vibrotactile stimuli with frequencies within the flutter range. Each stimulus pair (f_1, f_2; class) consisted of a large, unambiguous difference (8 Hz in most classes; Fig. 1B). In Fig. 1B we showed that performance for each class was only partially predictable from the absolute difference in magnitude, $\Delta_f = | f_1 - f_2 |$. A contraction bias could explain this disparity, since the decision accuracy depended on the specific stimuli that were paired. This internal process shifts the perceived frequency of the base stimulus f_1 towards the center of the stimuli frequency range ($<f_1>$, Fig. 1C). As such, if f_1 has a low stimulus value (i.e., 10 Hz), it will be perceived as larger, while a large frequency for f_1 is perceived as smaller. Therefore, correct evaluations for classes with $f_1 < f_2$ are increasingly obfuscated by the bias's

effect as classes are visited from right to left (upper diagonal, Fig. 1B-C), while classes with $f_1 > f_2$ are increasingly facilitated by the bias (lower diagonal, Fig. 1B-C). To preserve this bias structure, we labeled classes with beneficial bias effects as closer to the extreme ends, and classes with hindrance from the bias as the center classes (Fig. 1C). In this class organization, decision accuracy is highest for the classes near to the two ends (bias benefit; Fig. 1D), while their reaction time is shortest for correct trials (Fig. 1E). The decision accuracy (Fig. 1D) and RT (Fig. 1E) curves were U-shaped as a function of class number; center classes had the lowest accuracy and longest RTs (Methods). Opposingly, the monkey reacted slower for the extremum classes during error trials, faster for the center classes, and the result is an inverted U-shape curve (Fig. 1F).

Contraction bias affects DA responses

We recorded single-unit activity from 25 putative midbrain DA neurons while trained monkeys performed the discrimination task. They were identified based on previous electrophysiological criteria (Schultz and Romo, 1987): regular and low tonic firing rates (mean ± SD = 4.7 ± 1.4 spikes/s), and a long extracellular spike potential (2.4 ms ± 0.4 ms). Selecting neurons with positive phasic activation during reward delivery, and diminished phasic activity for error trials within the same period (Bromberg-Martin et al., 2010; Morris et al., 2006; Romo and Schultz, 1990), we calculated the response peaks for correct trials and plotted the values from all 25 neurons as a distribution (Fig. 2A, Methods). We observed a division within the population, allowing us to isolate two groups with a specific criterion (15Hz, dashed gray line). Hereafter, we labeled the subpopulation with lower peak responses as *"A"* (n = 12; Fig. 2A, orange) and with higher peak responses as *"B"* (n = 13; Fig. 2A, green). The majority of subgroup *A* (66%) showed significant, positive modulation to at least one of the two stimuli (Fig. 2B, orange

symbols), while the overwhelming majority of subgroup *B* did not respond significantly to either stimuli, excepting one neuron (Fig. 2B, green symbols). Based on these dynamic differences, we focus our study on subpopulation *A* from hereon.

In frontal lobe neurons, f_1 is parametrically represented during WM and the animal's decision is observable during the comparison period (Hernández et al., 2010). We hypothesized that if DA neurons receive feedback information from the frontal lobe, their responses to f_2 should reflect the contraction bias's effect as part of the decision report. Resultantly, neurons in subpopulation *A* showed significant dependence on stimulus classes (one-way ANOVA, p=0.0033; Fig. 2C). The U-shaped modulation in the response distributions can be explained by a pair of arguments: first, the bias induced a subjective difficulty that affected belief states. Second, as an RPE signal the DA response to f_2 coded the change in reward expectation produced by the application of the comparison frequency, which is higher for classes facilitated by the bias (Fig. 2C). A reinforcement learning model (Sutton and Barto, 2018) based on belief states supports this interpretation (Fig. S1A). Similar effects of belief states on responses to reward-predicting stimuli were observed in other experiments (Lak et al., 2017; Nomoto et al., 2010; Sarno et al., 2017). However, in those works the trial difficulty was introduced by experimental design while it was determined by a perceptual bias in our task.

Afterwards, we wondered whether the same bias effect was observable during f_1. By design, each f_1 value is the same for two classes ($f_1 > f_2$ or $f_1 < f_2$), where one class is favored by the bias, and the other is disfavored (Fig. 1D). For our query, we studied the monkeys' performance based on the f_1 value only. We observed that the decision accuracy was lowest for the extremum f_1 intensities (Fig. 2D, Right). As we would expect, the phasic DA response to f_1 was greatest for the middle values of our stimulus range (one-way ANOVA, p = 0.0034, Fig. 2D, Left).

Subpopulation *B* neurons were found to be invariant to class number (Fig. S2A) and f_1 (Fig. S2B), with small, unmodulated responses to f_2.

Contraction bias controls choice confidence

We reasoned that the contraction bias could influence the animal's decision confidence, similar to the effect based on stimulus features (Kepecs et al., 2008; Kiani and Shadlen, 2009). To answer this question, we used a Bayesian approach in which observation probabilities were combined with prior knowledge to obtain a posterior probability or belief about the state of the individual stimuli (Ashourian and Loewenstein, 2011; Jazayeri and Shadlen, 2015; Knill and Richards, 1996; Ma and Jazayeri, 2014; Preuschhof et al., 2010) (Fig. 3A). Choices were made according to the belief b_c, which corresponds to the largest of the two posterior probabilities: the belief about the state $f_1 > f_2$ (b(H), Higher) or the belief about $f_1 < f_2$ (b(L), Lower). The prior probabilities of f_1 and f_2 were assumed to be uniform, and we supposed that animals only perceived noisy representations of the two stimuli, despite having prior knowledge of the classes (Fig. 1B). We obtained observations of f_1 and f_2 from Gaussian distributions at the end of the delay period, with standard deviations σ_1 and σ_2 as noise parameters. We adjusted these parameters to optimize the similarity between the Bayesian model and the animal's performance (Fig. 3B, see Methods for model and parameter fitting). The best fit yielded $\sigma_2 = 3.2$ Hz and $\sigma_1 = 5.50$ Hz (Fig. 3B), so the noise parameter for f_1 was greater than that for f_2 ($\sigma_1 > \sigma_2$); from this, we intuited that the representation of f_1 deteriorated over the WM period. When this happens, the Bayesian model produces a contraction bias because the inferred value of f_1 is noisier than that of f_2, and the inferred value of f_1 relies more heavily on its prior distribution (Ashourian and Loewenstein, 2011).

Including these noise parameters, we used the Bayesian model to estimate confidence. We took confidence to be a measure of certainty that any given choice is correct. Within our Bayesian framework, confidence is given by b_c, i.e. the largest of the two beliefs, b(L) or b(H) (Fig. 3C). Using this model, we averaged over many simulated trials per class to analyze the level of confidence for each. In correct trials, confidence versus class number was U-Shaped (Fig. 3D, blue line), meaning confidence increased with the bias's favor. The two classes (11 and 12) with smaller absolute magnitude difference Δ_f (6Hz and 4Hz) had increased confidence due to bias, although this effect was negated by the increased difficulty in discrimination. Class 1 had the largest Δ_f (10Hz), and as such the bias increased confidence level while the difficulty of discrimination was less. Comparing the results between correct and error trials, the model had the opposite confidence pattern when they erred, demonstrating an inverted-U shape (Fig. 3C, red line). Both of these simulated results obtained from our Bayesian model parallel our psychophysical results. Thus, the contraction bias regulated confidence in a way similar to how an odor mixture modulated the choice confidence of rodents executing an odor categorization task (Kepecs et al., 2008). However, the important distinction for our results is that confidence was mainly modulated by an internally generated process, and not by the physical properties of the stimuli.

Confidence shapes DA responses during decision formation

To study the influence of confidence on DA activity, we subdivided our stimulus set (12 classes) into two groups: a high-confidence group (classes 1-3 & 10-12, Fig. 1C pink) and a low-confidence group (classes 4-9, Fig. 1C purple). During presentation of f_1, the two groups showed no significant difference in normalized activity (z-score, p = 0.58, two tail t-test; Fig. 4A, left; Methods). However, the two groups differed in temporal profiles during the presentation of f_2

(Fig. 4A, center), which we understood as exhibiting a disparity in confidence. We obtained a response latency of 250 ms for the time it took the confidence groups to significantly diverge in responses after f_2 onset (AUROC, $p<0.05$, sliding 250 ms window, 10 ms steps; Methods). Dependence on confidence during decision formation was also reported in previous works (Lak et al., 2017; Sarno et al., 2017). Interestingly, the activity after delivery of reward was invariant to confidence level, peaking at the same (Fig. 4A, right).

Our Bayesian model predicted lower confidence for error trials (Fig. 3D), so we studied the phasic activity of trials by decision outcome (correct v. incorrect). Responses to f_1 did not vary based on outcome (Fig. 4B, left), but the responses did vary during the comparison. The temporal profiles for each outcome differentiate with an analogous latency value (~260 ms) to the confidence groups (Fig. 4B, center). This time lag in DA outcome signals is comparable with those found in the secondary somatosensory cortex (Romo et al., 2002) and frontal areas during the same task (Hernández et al., 2010). These similarities in latency between frontal lobe and DA neurons were also found during a tactile detection task (de Lafuente and Romo, 2012), suggesting potential cooperation in higher order cortical functions.

At the point of reward delivery, the phasic DA activity increased in correct trials, but decreased for error trials (Fig. 4B, right). This temporal pattern is consistent with that predicted by the RPE hypothesis. The latency of divergence of these two response profiles (AUROC, $p <0.05$, Methods) was shorter than those after f_2 onset (~130 ms). Further, we asked whether our reinforcement learning model based on belief states could emulate the RPE dopamine responses. The model reproduced the phasic responses observed in Fig. 4 well (Fig. S1 B-C). As such, a greater influence from contraction bias correlates with lower accuracy and confidence, so the RPE was mediated by the subjective difficulty.

Task difficulty and the dispersion of the belief state

Bayesian decision theory provides us with an interpretation of the confidence and phasic DA response results we have presented. In correct trials, the difference in confidence level arises from the dispersion of the belief that any class k has been presented when the true class was k_0 (denoted as $B(k|k_0)$, Methods). In considering the most favored and disfavored classes on the upper diagonal (k_0 ={1,2} and k_0 ={5,6}, Fig. 5A), we summarized the results for all 12 classes. Note that orange circles in Fig. 5 indicate the actual class. Correct trials with low-confidence occur for classes disfavored by the bias (k_0 ={5,6}, Fig 5A), due to the spread of the belief $B(k|k_0)$ over classes k on the lower diagonal. In other words, confidence, the largest of the beliefs b(H) and b(L), is at its minimum (blue bar in the inset of Fig. 5A and blue line in Fig. 3D). The opposite is true for classes favored by the bias (k_0 ={1,2}, Fig. 5A): the belief is concentrated in a few possible options for class k, all of them on the upper diagonal. Lower belief (b(L) in the example classes) makes discrimination more difficult, so the reward expectation is lower; greater belief makes for an easier decision, and thus the reward expectation is larger. This mechanism produces DA responses (and RPEs) that reflect the confidence level in correct trials.

Pursuant to the same logic, error trials occur when the total belief about the wrong choice is greater than 0.5 (Fig. 5B; see red bars in insets). With a belief distributed almost evenly among correct and incorrect classes, confidence is diminished, explaining the lower activation of DA neurons in error trials (Fig. 4B, center). For classes favored by the contraction bias (k_0={1,2}), errors are likely due to noisy observations of f_2, which concentrates the belief state $B(k|k_0)$ in a few classes k located on the wrong side of the diagonal f_1=f_2 (e.g. k_0=1, Fig. 5B). In disfavored classes (k_0={5,6}, Fig. 5B), noisy observations of f_2 and the bias spread the belief $B(k|k_0)$ across several classes (k) with the wrong choice. The corresponding belief (b(H) in the example classes) assumes slightly higher values compared to those of the favored classes (k_0={1,2}, Fig. 5B),

meaning that during errors, the confidence in the incorrect choice results slightly higher (red bar in the inset of Fig. 5A and red line in Fig. 3D).

Do DA neurons code working memory?

Is the WM activity of DA neurons sustained and tuned to f_1? During the WM delay, neurons in frontal areas exhibit a temporally modulated sustained activity, tuned parametrically to f_1 (Brody et al., 2003; Hernández et al., 2010; Romo et al., 1999). The normalized activity (Methods) of an example neuron with significantly increased activity through the entire WM period is shown in Fig. 6A (left). When we sorted the firing rates by f_1 values, we saw a clear ramping in activity across the WM period in all 6 curves (Fig. 6A, center). However, the temporal profiles are not modulated by f_1 stimuli. To clarify these conflicting results, we averaged the entire WM activity for each f_1 stimulus (Fig. 6A, right). This neuron did not exhibit significant differences across f_1 (one-way ANOVA, $p = 0.69$). The second example neuron exhibited absolutely no temporal modulation during WM, remaining closer to its baseline value (Fig. 6B, left). When we expanded the observations to differentiate between responses to each f_1 stimulus, the ramping activity was still not observable (Fig. 6B, center). Further, neither the temporal profile nor the averaged activity during WM (Fig. 6B, center) were modulated based on f_1 (one-way ANOVA, $p = 0.81$).

At the population level, we found that neurons responded to f_1 by increasing their firing rate, but it was not tuned to the value of the frequency (one-tailed Wilcoxon signed-rank test; $p<0.001$; Fig. 6C). Furthermore, dividing the WM period into 3 segments (Fig. 6C, black horizontal bars indicated in the delay period), we observed positively and significantly modulated activity in all of them when compared to a segment before the presentation of f_1 (Fig. 6C, pink bar; one-tailed Wilcoxon signed-rank test; $p= 0.002$, $p<0.001$ and $p<0.001$, respectively). The average

population activity exhibited a ramping increase during the WM period (Fig. 6D). This behavior is not compatible with the usual reinforcement learning models (Schultz et al., 1997).

Due to the fact that we know that frontal neurons can be monotonically tuned to the identity of f_1 during WM, we sought to confirm that this coding did not exist within the DA population in any time interval. A linear regression in five discrete intervals (pink and black bars in Fig. 6C, top) showed no significant linear trend (Fig. 6E). Moreover, we extended our analysis to one second before f_1 onset through the end of the WM (a sliding 250 ms window, step 10 ms). In these 4.5 seconds, we performed linear and sigmoidal regressions in each window and found no significant results for either regression, in any window (slope different from zero, $p < 0.01$, for either a linear or sigmoidal fit with $Q > 0.05$; Methods). Notably, the absence of monotonic dependence held even during the last portion of the delay interval. This contrasts with the f_1 information recovery observed in frontal neurons (Romo and Rossi-Pool, 2020), a result also found in other tactile tasks (Rossi-Pool et al., 2016). Note that ramping activity in pure temporal frontal neurons can occur without coding the stimulus parameter in WM, similar to the neuron shown in Fig. 6A (Brody et al., 2003; Rossi-Pool et al., 2019).

Phasic and tonic DA activity depends on reaction time

We wanted to understand the trial-by-trial variance in RTs, so we classified trials according to RT into two groups (short & long, Methods). Dividing the reaction times into percentiles allowed us to observe a relationship between correct trials and reaction times: namely, longer reaction times tended to have less correct trials (Fig. S3A). We analyzed the time courses of DA activity for short- and long-RT trials (Fig. 7A), and we noticed that phasic responses to f_1 and f_2 correlated significantly with RTs, such that short-RT trials displayed greater activity (Fig. 7A). A similar dependence of phasic DA responses on RT has also been observed in a different task (Satoh et

al., 2003). Contrastingly, both groups reached a common activity peak response after reward delivery (Fig. 7A, right), but long-RT trials maintained a consistently higher activity for about 170 ms after reward delivery onset (AUROC, $p < 0.05$; Methods). This relationship between longer RTs and greater RPE signal can be understood in terms of reinforcement learning: longer RTs occurred mostly for classes negatively impacted by the contraction bias (Fig. 1E), so the network is reinforcing the positive evaluation since reward was unexpected (Fig. 7A, right). Contrastingly, the peak activity after reward was invariant to RT, which implies that the influence from class memory was lost. This is an intriguing result: on one hand, it suggests that a reset mechanism erased information about the trial type before the peak time but, on the other, such a mechanism would be expected to occur before the animal pressed a button to communicate its choice (Sarno et al., 2017). This effect was not significant when the response to reward delivery was sorted by class confidence levels (Fig. 4A, right); a plausible explanation is that sorting trials according to the average class confidence might have hidden the trial's perceived difficulty.

The DA signals during WM for short-RT trials were slightly, although consistently, greater than those found in long-RT trials. Although for most of that period these differences were not significantly different (Fig. 7A; the red bar indicates short periods of significant differences; AUROC, $p < 0.05$), once we averaged the activity for the RT groups across the whole WM period, the group activity of short-RT trials was significantly higher ($[\text{mean} \pm \text{SEM}]_{\text{short}}=0.25\pm0.03$; $[\text{mean} \pm \text{SEM}]_{\text{long}}=0.16\pm0.03$; $p=0.02$, two-sample one-tailed t-test). Mirroring subpopulation *A*, the activity of population *B* (Fig. 2A) was temporally modulated during WM as well, also distinguishing between short and long RTs. Additionally, the averaged WM activity for correct, short-RT trials was significantly higher than in long-RT trials ($[\text{mean} \pm \text{SEM}]_{\text{short}}=0.20 \pm 0.06$; $[\text{mean} \pm \text{SEM}]_{\text{long}}=0.07\pm0.04$; $p=0.03$, two-sample one-tailed t-test).

Reward expectation on short- and long-RT trials

Dopamine activity responded with different time courses for trials with distinct RTs, so we tested for any f_1 dependencies during WM for each of the RT trial groups. We focus on the time period from 0.5 seconds before f_1 onset through f_2 onset (WM delay period). Sorting each RT trial group by the values of f_1, we searched for the windows with significant ANOVA results ($p<0.05$; Methods). Consequently, we identified many intermittent instances during f_1 presentation and throughout WM with significant f_1 dependence. This led us to seek temporal consistency, so we divided the entire time range into 8 discrete subintervals of 500 ms and calculated the fraction of significant sliding windows in each (Methods). Trials with long RTs showed no dependence on f_1 throughout the time range, but short-RT trials had consistent dependence on f_1 for about half of the associated presentation period, as well as the first 500 ms of WM. Two examples of short-RT responses during the presentation of f_1 (t=0.225s) and during the first 500 ms of the delay period (t=0.715s) are exhibited in Fig. 7B. The corresponding z-scores depended on f_1 in a similar way as the mean DA activities shown in Fig. 2D. During the rest of the delay period short-RT trials did not depend significantly on f_1.

After studying f_1 in isolation, we considered dependencies on stimulus class. Using the two RT trial groups sorted by stimulus class, we analyzed the period between f_2 onset and 1 second after f_2 offset (1.5 second interval). We implemented the ANOVA test and searched for those windows in which $p<0.05$. As before, we calculated the temporal consistency in 3 discrete subintervals of 500 ms, based on the fraction of significant 250 ms windows; short-RT trials had significant class-dependent activity during f_2 stimulation (t=3.685), with a pattern similar to that in Fig. 2C (Fig. 7C, left). On the other hand, long-RT trials were significantly temporally consistent in all three subintervals, with a strong class dependency during f_2 (t=3.655s, Fig. 7C,

right). Intriguingly, the activity pattern as a function of the class was far more complex in the subinterval after f_2 offset (t=4.505s, Fig. 7D, left). A visualized interpretation was attained when we plotted the phasic activity by the base frequency f_1 (t=4.505s, Fig. 7D, right): the U-shaped pattern is a clear inversion of the phasic response to f_1 of the DA neurons in short-RT trials (compared with Fig. 7B). This means that this response for long-RT trials was greater for the classes that were affected in a stronger manner by the contraction bias, a result consistent with a role of phasic DA activity as a RPE signal before pressing the button. If the expected total reward after the application of f_1 is $Q(f_1)$, the RPE can be estimated as the expected reward for pressing the chosen button (which is independent of f_1) minus $Q(f_1)$. This subtraction leads to the inversion of the f_1 dependence observed in Fig. 7D (right) with respect to Fig. 7B. This inversion at the go-cue event was also observed in a detection task (Sarno et al., 2017). This effect was not observed in short-RT trials, plausibly because when the animal's reaction was fast this uncertainty dependence was more difficult to record.

DA stimulus responses reflect motivated behavior on a trial-by-trial basis

The results presented above suggest that the phasic DA activity codes motivation in its response (Fig. 7A). In a task without working memory (Satoh et al., 2003), it was found that the motivational level was negatively correlated with the phasic DA response associated to a reward-predicting cue on a trial basis. In our task, the evidence used to reach a decision is uncertain, so when we present class k to the monkey, based on its beliefs, it could have inferred that the class was k_0 (Fig. 5). This means that the animal's perceived difficulty and its RT were not associated to the true class k but to class k_0. Then, to estimate correlations at fixed perceived difficulty (class k_0) between RT and phasic DA responses, one should employ trials from all the classes k, each weighted with the corresponding belief, $B(k_0|k)$ (Methods). The RTs and these weighted DA

responses were negatively correlated (Fig. 8A). This means that DA signals carry motivational information during the comparison period. The correlation was modulated by class number: while greater bias obstruction gives rise to a strongly negative correlation coefficient (-0.25), classes favored by the bias were almost not correlated at all (~0). These results indicate that DA responses during f_2 were related to motivation, on a trial-by-trial basis, in the greatest manner for the most difficult trials. Furthermore, DA responses to f_1 were also correlated with RT. The association became more negative as the f_1 frequency intensified (Fig. 8B). This dependence supports DA as a driver for motivation based on trial difficulty. Considering Weber's law, frequency discriminability is facilitated by lower f_1 values, which makes low-f_1 trials be perceived as easier to handle. This association between ease of f_1 frequency, ease of stimulus class, and RTs is further explored in Fig. S3B-C.

Our results have shown that the DA activity during the WM was not tuned to f_1, did not reflect reward-related information, although it did predict RTs. To estimate how the dopamine RT correlation changed across the WM period would require additional hypotheses about the changing representation of f_1. To avoid that estimation, and to obtain qualitative insight into the DA engagement in motivated behavior, we considered the correlation between the mean DA activity during WM and the RT, weighted with the initial belief about f_1 (Methods). The dependence of this correlation on f_1 had a similar trend to that during the application of f_1 for short-RT trials. Further study would inform if this result actually demonstrates the motivation's continuous influence during WM up to the reception of the comparison stimulus (f_2).

DISCUSSION

We sought to determine the activity of DA neurons recorded while monkeys performed a cognitively demanding discrimination task. We found that phasic DA responses coded RPE and

motivation, although the dopamine activity during WM only coded motivation. We also observed that while in frontal cortex a neural representation of f_1 is maintained throughout the WM delay with a sustained activity pattern (Brody et al., 2003; Hernández et al., 2010; Romo et al., 1999), a subset of DA neurons exhibited a ramping increase during the delay.

The contraction bias affected the task difficulty. Naturally occurring signals vary slowly across time, and the brain has adapted circuits able to anticipate future stimuli based on an average of recently past sensory inputs (Dyjas et al., 2012; Raviv et al., 2012). In our experiment, we applied frequencies in a pseudorandom order, producing abrupt changes in frequency between consecutive trials. The contraction bias arises from the contradiction between the expected and received stimuli, influencing the subjective task difficulty. Our results suggest that DA neurons process reward-related information with respect to signals received from those adapted circuits.

A model-based approach allowed us to estimate the RPE signal in simulated DA neurons using belief states (Gershman and Uchida, 2019; Lak et al., 2017; Rao, 2010; Sarno et al., 2017; Starkweather et al., 2017); our model showed that the contraction bias controlled choice and confidence in a significant manner, as well as modulating the phasic DA activity during presentation of f_2. The responses during reward were initially affected by the bias-induced subjective task difficulty, although the peak of response was invariant to stimulus class. Contrastingly, DA neurons studied in the random dots motion paradigm showed that the peak response to a feedback tone, announcing the trial outcome, did vary based on motion coherence (Lak et al., 2017); these discrepancies between studies may be explained by distinct experimental designs.

The phasic responses to f_1 and f_2 depended on RT, increasing for faster responses. Fast DA fluctuations (phasic responses) have been proposed as potential motivational drivers (Berke,

2018; Satoh et al., 2003). In this respect, we found that in our task phasic DA responses to uncertain sensory signals were correlated with RTs on a trial basis, with strength modulated by perception of task difficulty throughout a trial. At f_1 presentation, the activity correlation with RTs was stronger for those f_1 frequencies that were most challenging for discrimination. When f_2 was applied, the bias obstructed class identification, and the negative DA-RT correlation resulted stronger for classes disfavored by the bias.

In addition to motivation dependence, the contraction bias produced a phasic dependence on f_1 that coded reward expectation in short-RT trials. This effect abated for long-RT trials, possibly explained by lower motivation obscuring representation of task-relevant cues (Etzel et al., 2016; Kennerley and Wallis, 2009). In contrast, the f_2 associated response correlated with class number in all trials, regardless of RT. We consider that this could arise from f_2 completing the stimulus class information, so the evaluation becomes dominant in producing the response. An unexpected result was that DA neurons exhibited RT-dependent sustained activity during WM. This DA signal can convey motivational information, which could be playing many roles in cortical cognitive processes. An interesting hypothesis relates motivation with cognitive control modulated by dopaminergic signals to prefrontal cortex and striatal neurons (Botvinick and Braver, 2015; Cools, 2016; Yee and Braver, 2018): DA is proposed as a persistent-activity stabilizer, increasing the gain in target neurons, and in turn producing an enhanced signal-to-noise ratio and promoting WM stability (Brunel and Wang, 2001; Durstewitz and Seamans, 2002; Servan-Schreiber et al., 1990).

Previous recordings in midbrain DA neurons during monkeys' performance of WM tasks did not observe sustained activity during the delay (Ljungberg et al., 1991; Matsumoto and Takada, 2013; Schultz et al., 1993). This discrepancy can be attributed to difference in WM and attention

requirements, as well as to task difficulty. On the other hand, to our knowledge, the ramping WM activity we observed has not been reported before for these neurons. In the ventral striatum, however, a ramping-increase in release of DA has been described (Berke, 2018; Hamid et al., 2016; Howe et al., 2013; Roitman et al., 2004); this signal purportedly reflects an approaching decision report (Howe et al., 2013), possibly aiding in action invigoration as well (Westbrook and Braver, 2016). Several explanations have been proposed for generating this ramping signal (Gershman, 2014; Kato and Morita, 2016; Lloyd and Dayan, 2015). A reinforcement learning framework that includes the selection of the action and of when and how fast to act (vigor), gives rise to ramping signals (Lloyd and Dayan, 2015; Niv et al., 2007). In our experiment, we were not able to address the implications of the WM ramping signal. However, we believe that this DA ramping signal is not associated to the physical effort; the initiation of a decision report should occur near PB, several seconds after the end of the WM period. Instead, we propose that this activity is due to the cognitive effort related to upload and maintaining a percept in WM.

Recent advances in the study of motivation and cognitive control (Yee and Braver, 2018) indicate possible directions to investigate this issue further in future work. Cognitive control is effortful (Kool and Botvinick, 2013; Shenhav et al., 2013), involving a cost-benefit decision-making process. A task that involves WM requires an evaluative choice: either engaging in the task, or choose a less demanding option (e.g., guessing), even if it leads to a smaller reward (Manohar et al., 2015; Westbrook and Braver, 2016). The cost-benefit calculations include the intensity of the cognitive action, which determines the intrinsic cost of cognitive control. A normative approach combining the expected reward, the intrinsic cost of control and the cost of vigor could generate a rising delay-period DA activity.

Summarily, our results underline the essential role of midbrain DA in learning, motivation and working memory under perceptual uncertainty. In our frequency discrimination task, phasic DA responses were affected by the contraction bias and motivation, while tonic activity could instead be compatible with cognitive control. The dependence of DA activity on the contraction bias suggests that internally generated biases can play role in learning. Two-interval forced-choice tasks, such as temporal interval discrimination, have studied behavior, contraction bias and stimulus sequence effects (Dyjas et al., 2012); however, DA was not investigated in these cases. Future study within these task paradigms can further our understanding of DA activity in motivation, cognitive control, and during working, expanding on the influence of internal biases on evaluation and learning.

ETHICS

Experimental procedures were approved by the Institutional Animal Care and Use Committee of the Cell Physiology Institute, National Autonomous University of Mexico (protocol number 046)

ACKNOWLEDGMENTS

This work was supported by Grants PGC2018-101992-B-I00 from the Spanish Ministry of Science, Innovation and Universities (to S.S., M.B. and N.P.) and by the Dirección de Asuntos del Personal Académico de la Universidad Nacional Autónoma de México, Consejo Nacional de Ciencia y Tecnología PAPIIT-IN210819 (to R.R. and R.R.-P.).

AUTHOR CONTRIBUTIONS

R.R., G.G.-d.L. and R.R.-P. performed experiments. S.S., M.B. and N.P. analyzed the data and designed the models. S.S. and M.B. wrote the codes. N.P., S.S, R.R-P and R.R. wrote the paper helped by the other authors.

REFERENCES

Akrami, A., Kopec, C.D., Diamond, M.E., and Brody, C.D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. Nature *554*, 368–372.

Ashourian, P., and Loewenstein, Y. (2011). Bayesian Inference Underlies the Contraction Bias in Delayed Comparison Tasks. PLoS One *6*, e19551.

Babayan, B.M., Uchida, N., and Gershman, S.J. (2018). Belief state representation in the dopamine system. Nat. Commun. *9*, 1–10.

Berke, J.D. (2018). What does dopamine mean? Nat. Neurosci. *21*, 787–793.

Botvinick, M.M., and Braver, T.S. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. Annu. Rev. Psychol. *66*, 83–113.

Brody, C.D., Hernández, A., Zainos, A., and Romo, R. (2003). Timing and Neural Encoding of Somatosensory Parametric Working Memory in Macaque Prefrontal Cortex. Cereb. Cortex *13*, 1196–1207.

Bromberg-Martin, E.S., Matsumoto, M., and Hikosaka, O. (2010). Distinct Tonic and Phasic Anticipatory Activity in Lateral Habenula and Dopamine Neurons. Neuron *67*, 144–155.

Brunel, N., and Wang, X.J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J. Comput. Neurosci. *11*, 63–85.

Cools, R. (2016). The costs and benefits of brain dopamine for cognitive control. Wiley Interdiscip. Rev. Cogn. Sci. *7*, 317–329.

Durstewitz, D., and Seamans, J.K. (2002). The computational role of dopamine D1 receptors in working memory. Neural Networks *15*, 561–572.

Dyjas, O., Bausenhart, K.M., and Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. Attention, Perception, Psychophys. *74*, 1819–1841.

Etzel, J.A., Cole, M.W., Zacks, J.M., Kay, K.N., and Braver, T.S. (2016). Reward Motivation Enhances Task Coding in Frontoparietal Cortex. Cereb. Cortex *26*, 1647–1659.

Fassihi, A., Akrami, A., Esmaeili, V., and Diamond, M.E. (2014). Tactile perception and working memory in rats and humans. Proc. Natl. Acad. Sci. *111*, 2331 LP – 2336.

Gershman, S.J. (2014). Dopamine ramps are a consequence of reward prediction errors. Neural

Comput. *26*, 467–471.

Gershman, S.J., and Uchida, N. (2019). Believing in dopamine. Nat. Rev. Neurosci. *20*, 703–714.

Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., and Berke, J.D. (2016). Mesolimbic dopamine signals the value of work. Nat. Neurosci. *19*, 117–126.

Hernández, A., Salinas, E., García, R., and Romo, R. (1997). Discrimination in the Sense of Flutter: New Psychophysical Measurements in Monkeys. J. Neurosci. *17*, 6391–6400.

Hernández, A., Nácher, V., Luna, R., Zainos, A., Lemus, L., Alvarez, M., Vázquez, Y., Camarillo, L., and Romo, R. (2010). Decoding a Perceptual Decision Process across Cortex. Neuron *66*, 300–314.

Hollingworth, H.L. (1910). The Central Tendency of Judgment. J. Philos. Psychol. Sci. Methods *7*, 461–469.

Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E.M., and Graybiel, A.M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature *500*, 575–579.

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

Kato, A., and Morita, K. (2016). Forgetting in Reinforcement Learning Links Sustained Dopamine Signals to Motivation. PLoS Comput. Biol. *12*, e1005145.

Kennerley, S.W., and Wallis, J.D. (2009). Reward-Dependent Modulation of Working Memory in Lateral Prefrontal Cortex. J. Neurosci. *29*, 3259–3270.

Kepecs, A., Uchida, N., Zariwala, H.A., and Mainen, Z.F. (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature *455*, 227–231.

Kiani, R., and Shadlen, M.N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. Science (80-.). *324*, 759–764.

Knill, D.C., and Richards, W. (1996). Perception as Bayesian Inference (Cambridge: Cambridge University Press).

Kool, W., and Botvinick, M.M. (2013). The intrinsic cost of cognitive control. Behav. Brain Sci. *36*, 697–698.

Körding, K. (2007). Decision Theory: What "Should" the Nervous System Do? Science (80-.). *318*, 606–610.

de Lafuente, V., and Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. Proc. Natl. Acad. Sci. U. S. A. *108*, 19767–19771.

de Lafuente, V., and Romo, R. (2012). Dopaminergic activity coincides with stimulus detection

by the frontal lobe. Neuroscience *218*, 181–184.

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., and Kepecs, A. (2017). Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. Curr. Biol. *27*, 821–832.

Ljungberg, T., Apicella, P., and Schultz, W. (1991). Responses of monkey midbrain dopamine neurons during delayed alternation performance. Brain Res. *567*, 337–341.

Lloyd, K., and Dayan, P. (2015). Tamping Ramping: Algorithmic, Implementational, and Computational Explanations of Phasic Dopamine Signals in the Accumbens. PLoS Comput. Biol. *11*, e1004622.

Ma, W.J., and Jazayeri, M. (2014). Neural Coding of Uncertainty and Probability. Annu. Rev. Neurosci. *37*, 205–220.

Manohar, S.G., Chong, T.T.-J., Apps, M.A.J., Batla, A., Stamelou, M., Jarman, P.R., Bhatia, K.P., and Husain, M. (2015). Reward Pays the Cost of Noise Reduction in Motor and Cognitive Control. Curr. Biol. *25*, 1707–1716.

Matsumoto, M., and Takada, M. (2013). Distinct Representations of Cognitive and Motivational Signals in Midbrain Dopamine Neurons. Neuron *79*, 1011–1024.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. *9*, 1057–1063.

Niv, Y., Daw, N.D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl). *191*, 507–520.

Nomoto, K., Schultz, W., Watanabe, T., and Sakagami, M. (2010). Temporally Extended Dopamine Responses to Perceptually Demanding Reward-Predictive Stimuli. J. Neurosci. *30*, 10692–10702.

Preuschhof, C., Schubert, T., Villringer, A., and Heekeren, H.R. (2010). Prior Information Biases Stimulus Representations during Vibrotactile Decision Making. J. Cogn. Neurosci. *22*, 875–887.

Rao, R. (2010). Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. Front. Comput. Neurosci. *4*, 146.

Raviv, O., Ahissar, M., and Loewenstein, Y. (2012). How Recent History Affects Perception: The Normative Approach and Its Heuristic Approximation. PLOS Comput. Biol. *8*, e1002731.

Roitman, M.F., Stuber, G.D., Phillips, P.E.M., Wightman, R.M., and Carelli, R.M. (2004). Dopamine Operates as a Subsecond Modulator of Food Seeking. J. Neurosci. *24*, 1265–1271.

Romo, R., and Rossi-Pool, R. (2020). Turning Touch into Perception. Neuron *105*, 16–33.

Romo, R., and Schultz, W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J. Neurophysiol. *63*, 592–606.

Romo, R., Brody, C.D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric

working memory in the prefrontal cortex. Nature *399*, 470–473.

Romo, R., Hernández, A., Zainos, A., Lemus, L., and Brody, C.D. (2002). Neuronal correlates of decision-making in secondary somatosensory cortex. Nat. Neurosci. *5*, 1217–1225.

Rossi-Pool, R., Salinas, E., Zainos, A., Alvarez, M., Vergara, J., Parga, N., and Romo, R. (2016). Emergence of an abstract categorical code enabling the discrimination of temporally structured tactile stimuli. Proc. Natl. Acad. Sci. U. S. A. *113*, E7966–E7975.

Rossi-Pool, R., Zizumbo, J., Alvarez, M., Vergara, J., Zainos, A., and Romo, R. (2019). Temporal signals underlying a cognitive process in the dorsal premotor cortex. Proc. Natl. Acad. Sci. U. S. A. *116*, 7523–7532.

Sarno, S., de Lafuente, V., Romo, R., and Parga, N. (2017). Dopamine reward prediction error signal codes the temporal evaluation of a perceptual decision report. Proc. Natl. Acad. Sci. *114*, E10494–E10503.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated Coding of Motivation and Outcome of Decision by Dopamine Neurons. J. Neurosci. *23*, 9913–9923.

Schultz, W., and Romo, R. (1987). Responses of nigrostriatal dopamine neurons to high-intensity somatosensory stimulation in the anesthetized monkey. J. Neurophysiol. *57*, 201–217.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli during Successive Steps of Learning a Delayed Response Task. J. Neurosci. *13*, 900–913.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. Science (80-.). *275*, 1593–1599.

Servan-Schreiber, D., Printz, H., and Cohen, J.D. (1990). A Network Model of Catecholamiine Effects: Gain, Signal-to-Noise Ratio, and Behavior. Science (80-.). *249*, 892–895.

Shenhav, A., Botvinick, M.M., and Cohen, J.D. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. Neuron *79*, 217–240.

Starkweather, C.K., Babayan, B.M., Uchida, N., and Gershman, S.J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. Nat. Neurosci. *20*, 581–589.

Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (Cambridge, MA, USA: MIT Press).

Westbrook, A., and Braver, T.S. (2016). Dopamine Does Double Duty in Motivating Cognitive Effort. Neuron *89*, 695–710.

Yee, D.M., and Braver, T.S. (2018). Interactions of motivation and cognitive control. Curr. Opin. Behav. Sci. *19*, 83–90.

METHODS

Discrimination Task

This study was performed on two male monkeys, *Macaca mulatta*, 5–7 kg. The sensory discrimination task used here has been described previously (Romo and Rossi-Pool, 2020; Romo et al., 1999, 2002). The schematic representation of this task is depicted in Fig. 1A. The monkey sat on a primate chair with its head fixed. The right hand was restricted through a half-cast and kept in palm-up position. The left hand operated an immovable key (elbow at ~90°) and two push buttons in front of the animal, 25 cm away from the shoulder at eye level. The centers of the switches were located 7 and 10.5 cm to the left of the midsagittal plane. In all trials, the monkey first placed the left hand on the key, and later projected to one of the two switches. A variable delay of 1.5-3 s is presented to avoid anticipatory activity before the arrival of the stimulus, followed by the first stimulus (f_1), lasting 0.5 s. The second stimulus (f_2) is presented after a 3 s delay, also lasting 0.5 s. The offset of f_2 signals the monkey to release the key (Key Up, KU), and report its decision by pressing one of two push buttons (PB) with the left hand (lateral push button for $f_2 > f_1$, medial push button for $f_2 < f_1$). The time lag between the offset of f_2 and KU constitutes the reaction time (RT, Fig. 1A). Immediately after the decision report, correct discriminations were rewarded with a few drops of liquid, while incorrect discriminations received a few seconds of delay before the beginning of the next trial. Stimuli were delivered to the skin of the distal segment of one digit of the restrained right hand, via a computer-controlled stimulator (BME Systems; 2 mm round tip). Initial probe indentation was 500 μm. Vibrotactile stimuli were mechanical sinusoids pulses lasting 20ms each. Stimulation amplitudes were adjusted to produce equal subjective intensities (Hernández et al., 1997). Performance was quantified through psychometric techniques (Fig. 1B and D). Animals were handled in accordance with standards of the National Institutes of Health and Society for Neuroscience. All protocols were approved by the Institutional Animal Care and Use Committee of the Instituto de

Fisiología Celular (UNAM).

Recordings

Recordings were obtained with quartz-coated platinum-tungsten microelectrodes (2 to 3 MΩ; Thomas Recording) inserted through a recording chamber located over the central sulcus, parallel to the midline. Midbrain DA neurons were identified on the basis of their characteristic regular and low tonic firing rates (1-10 spikes per second) and by their long extracellular spike potential (2.4 ms ± 0.4 SD).

Neuron selection

From the total number of recorded neurons (n=196), we selected those cells with responses to reward delivery (correct trials) that were significantly higher than responses to lack-of-reward (miss trials; $P < 0.05$, two-sample t test). There were 25 recorded neurons that coincided with this criterion (n=25). For each neuron, we obtained the maximal response to reward delivery. The time of maximal response was assessed by computing the firing rate as a function of time, using 100 ms sliding windows displaced every 100 ms. The activity after the reward delivery was calculated for correct and error trials using a window of 200 ms and 500 ms, respectively, such that each was centered at the time of maximum response.

To further evaluate the selection criterion, we took the 25 selected neurons and shuffled the labels among correct and incorrect trials for each neuron (Romo et al., 1999). We used a permutation test with 1000 resamples to re-analyze the response of these neurons to reward delivery. Out of the 25 neurons, none were found to be responsive to the reward during shuffled instances. Thus, the net probability of marking a neuron as reward-responsive by chance was $P<1/25=0.04$.

To evaluate whether neurons responded positively to the stimuli (Fig. 2B) we performed a Wilcoxon signed rank test between the mean firing rate before stimulus presentation (f_1 and f_2) and during the stimulation period. For both time intervals the mean firing rate before the stimulus presentation was calculated in a 500 ms window centered 650 ms before stimuli onset. The mean firing rates during f_1 and f_2 were measured, respectively, in a 500 ms and 350 ms window centered 250 ms after the stimulus onset. To improve visualization, data points in Fig. 2B represent the mean firing rate during f_1 and f_2 standardized (z-score) with respect to a 500 ms window centered 650 ms before f_1.

Analysis of behavioral data

Behavioral data were obtained on average from 2226 trials per stimulus class (Fig. 1D). For each session, the reaction times (RTs) were standardized (z-score) with respect to the mean and STD for each of the two decision outcomes (Fig. 1E-F).

Analysis of the firing rate activity

The responses (z-scores) to f_1 and f_2 in Fig. 2C-D were standardized with respect to a temporal window preceding the onset of the base stimulus that lasted 500 ms and was centered 1000 ms after KD. To estimate the temporal profile of the z-score (Fig. 4, Fig. 6A-B left and center, Fig. 6C, and Fig. 7A) we calculated the firing rate for each neuron in 250 ms sliding windows shifted every 50 ms and standardized as in Fig. 2C-D. Finally, the responses (z-scores) during the delay period (Fig. 6A-B right) were measured during its entire duration (3 s) and standardized as in Fig. 2C-D.

Latency values

The firing rate time-course of the responses to f_2 depend on trial confidence and trial outcome (Fig. 4). To determine the time of divergence between the two time-courses we applied a receiver operating characteristic (ROC) curve analysis in each sliding temporal windows. This was done within the period lasting from 300 ms before f_2 onset to 200 ms after its offset. For each neuron we obtained the normalized firing rate (z-score) in sliding windows of 250 ms shifted in 10 ms steps. We used the z-scores of all neurons and trials to calculate the ROC curve at each time bin. The area under the ROC curve (AUROC) was used as the index indicating differential neuronal activity across different trial types. Values of the AUROC higher or lower than 0.5 indicated that, at the population level, one type of trials evoked a higher or lower DA response than the other. To determine the statistical significance of the computed AUROCs, we used a permutation test with 1000 resamples. Significance was determined with $p<0.05$ in 5 consecutive windows and the latency was defined as the first window that met this criterion. A similar analysis was used to determine significant differences in the temporal profile of the normalized activity for trials sorted according to the RT. Short-RT trials were defined as trials with RT below the 40th percentile and long-RT trials were those with RT above the 60th percentile (Fig. 7A).

Dependence of the DA activity on f_1 and class number

In order to search for f_1-dependent activity we performed two different tests. We first used a linear and a sigmoidal regression analysis (Fig. 6E) to assess whether the activity was monotonic with respect to f_1. Then, we used a one-way analysis of variance (ANOVA) test to identify any general, non-monotonic relationship between firing rate activity and f_1. We focused both analyses on the f_1-stimulation period and WM delay between f_1 and f_2. We calculated a mean time-dependent z-score (standardized as in Fig. 2C-D) using a sliding window of 250 ms moving in steps of 10 ms, from 0.5 s before f_1 onset up to f_2 onset. Window times with a significant

monotonic signal (slope different from zero, $p < 0.01$, for either a linear or sigmoidal fit with $Q > 0.05$) were marked as "significantly monotonic." Window times where the ANOVA was significant ($p < 0.05$) were marked as "significantly dependent" (Brody et al., 2003; Romo et al., 1999). We then divided the 4 s period (from 0.5 s before f_1 onset up to f_2 onset) into 8 non-overlapping intervals of 500 ms and counted the number of windows that were significantly linear in each interval. For each interval, we said that the f_1 dependence was significantly linear if more than the 40% of windows were significantly linear, and significantly dependent if more than 40% of windows gave a significant ANOVA p-value. Figs. 7B and D show the z-scores in windows in which the dependence was significant, respectively for short- and long-RT trials. A similar procedure based on the ANOVA test was employed to calculate how the responses during and after f_2 depended on the class number for both long- and short-RT trials (see Fig. 7C and D).

Bayesian model for the discrimination task

The discrimination task was modeled using a Bayesian framework. The prior probabilities of f_1 and f_2 were taken to be uniform and denoted as $P(f_1^i)$ (i = 1, …, 6) and $P(f_2^j)$ (j = 1, …, 10), respectively. It was assumed that the animal knew the class structure used in the experiment (Fig. 1B-C) but it had access only to noisy representations (observations) of the two frequencies presented in the trial (denoted by $f_{1,0}$ and $f_{2,0}$). At the onset of the second stimulus, an observation o_2 of the frequency $f_{2,0}$ was obtained from a Gaussian distribution with mean $f_{2,0}$ and standard deviation σ_2. This noisy information was combined with knowledge of the prior distribution $P(f_2^j)$ to obtain the belief, or posterior distribution about the value of the second frequency, $b_2(f_2^j) = P(f_2^j|o_2) \propto P(o_2| f_2^j)\, P(f_2^j)$. The observation of the first frequency $f_{1,0}$, made at the end of the delay period, had to be retrieved from working memory and was indicated by o_1^*. This

observation was also taken from a Gaussian distribution, but with mean $f_{1,0}$ and standard deviation σ_1. The belief about the value of the first frequency was denoted by $b_1(f_1^i) = P(f_1^i | o_1^*)$.

The belief state $B(k | o_1^*, o_2)$ about the class $c_k = (f_1^k, f_2^k)$ (k=1, …, 12; class labels, Fig. 1C) was defined as the set of the posterior probabilities $P(f_1 = f_1^k, f_2 = f_2^k | o_1^*, o_2)$ that the class c_k had been presented in this trial, conditioned to the observations o_2 and o_1^*. It can be written as:

$$B(k|o_1{}^*, o_2) = \frac{P(o_2)}{P(o_2|o_1*)} \frac{P(f_2=f_2^k|f_1=f_1^k)}{P(f_2^k)} b(f_1^k) b(f_2^k) \quad (1)$$

The first factor in the above equation is a normalization factor. The second factor is a transition matrix relating the first and the second stimulation frequencies divided by the prior probability of the second stimulation frequency. Since we assumed that the animal had perfect knowledge of the class structure, the only non-zero matrix elements correspond to the 12 classes c_k of the experiment. Furthermore, since f_2 can only take two possible values for a given value of f_1, all non-zero transition probabilities are 0.5. The last two factors then become the beliefs $b_1(f_1^k)$ and $b_2(f_2^k)$ about the first and second frequencies being those in class c_k.

The sums of the $B(k|o_1^*,o_2)$'s over classes k for each choice, gives the two belief values ($f_1 > f_2$ or $f_1 < f_2$). These two sums were denoted by b(H) (higher choice, $f_1 > f_2$) and b(L) = 1 − b(H) (lower choice, $f_1 < f_2$), and choices were made according to the larger of these two, denoted as b_c. The performance can be measured by the fraction of trials of a given class in which the decision is correct.

The two unknown model parameters, the standard deviations σ_1 and σ_2, were fitted by minimizing the mean squared error between the model performance and the animal's performance. We found the optimal parameter values: σ_1 =5.5Hz and σ_2 =3.2Hz.

The confidence C of a given trial is C=max[b(H), b(L)]=b_c, which is bounded between 0.5 and 1. The confidence of a given class was defined as the average of the value C across trials of each class. The confidence in hits (or errors) of a given class was defined as the average over the correct (or wrong) trials from each class.

Finally, the *class belief state* $B(k\,|k_0)$ was defined as the average of the belief state $B(k\,|o_1^*, o_2)$ over all trials from each class $c_{k0} = (f_{1,0}\,, f_{2,0})$.

Reinforcement Learning Model based on Belief States

To test whether the phasic responses can be attributed to dopamine reward prediction errors (RPEs) we constructed a reinforcement learning model and checked if it was able to reproduce the observed responses. Given that in the task the relevant stimuli are only partially observable (the animal is not aware of the true value of f_1 and f_2) we used a belief-state TD model (similar to that proposed by (Rao, 2010)) to compute reward expectations and simulate the RPEs signaling. This was first implemented in a Bayesian module that works similarly to the Bayesian model described above and uses the same fitted values of the two noise parameters, σ_1 and σ_2. This module yields the belief $b_1(f_1^i)$ (i=1, …, 6) about the value of the first frequency, the belief state $B(k\,|\,o_1^*, o_2)$ about the class $c_k = (f_1^k\,, f_2^k)$ (k=1, …, 12) at the time when the second frequency is presented and the belief b_c = [b(H), 1-b(H)] about which of the two frequencies is higher. Then it transmits these inference results to a temporal difference (TD) module that selects actions and generates reward prediction errors (RPEs).

The RL model also uses a fully observable variable *pm* (push movement) that represents the movement towards one of the two buttons. This variable has two possible states: *py* ("push yes") when the decision is $f_1>f_2$ and *pn* ("push no") when the decision is $f_1<f_2$. Finally, the reward

function, denoted by r, is a scalar function that takes two different values for correct and incorrect decisions (see Equation 8).

At the beginning of each trial the TD module calculates the value of the first stimulus as

$$V_1(b_1) = \sum_{i=1}^{6} Q_1(i) \cdot b_1(i) \tag{2}$$

where Q_1(i) is a set of adaptable weights. The RPE at the first stimulus is $\delta(f_1) = V_1(b_1)$.

At the second stimulus the TD module computes the value of the class:

$$V_B(B) = \sum_{k=1}^{12} Q_B(k) \cdot B(k|o_1{}^*, o_2). \tag{3}$$

Here the Q_B(k) (k=1, …,12) are another set of adaptable weights. At the second stimulus the TD module also estimates the value resulting from the comparison of the two frequencies

$$V(b_c) = \sum_{i}^{G} g_i(b_c) * v_i. \tag{4}$$

where the v_i are a set of adaptable weights. The g_i (i=1, …, G) are convenient functions that account for the contribution of the belief $b_c$to this value. The functions g_i were taken as the following basis functions

$$g_i(b_c) = \left(\frac{1}{2}\right)[cos(\mathrm{a}\,(b_c - \mathrm{c}_i)) + 1], \tag{5}$$

if $\mathrm{c}_i - 0.1 < b_c < \mathrm{c}_i + 0.1$ and $g_i(b_c) = 0$ otherwise. The c_i's are G=11 equally spaced centroids in [0,1] and a = π /0.1.

Given the values $V_B(B)$ and $V(b_c)$ the RPE at the second stimulus is

$$\delta(f_2) = V_B(B) + V(b_c) - V_1(b_1). \tag{6}$$

As in the Bayesian model, we assumed that decisions were made according to the larger of the beliefs b(H) and b(L) about which of the two frequencies was the higher.

The value of the response movement when action j is selected is indicated with V_{rm}(j) to highlight the correspondence between the action selected and the subsequent movement.

The RPE at the response movement is

$$\delta(rm) = V_{rm}(j) - V_B(B) - V(b_c) \tag{7}$$

and the RPE at the delivery of reward is

$$\delta(r) = r - V_{rm}(j), \tag{8}$$

where r =1 for correct discrimination and r = -0.5 otherwise.

The RPEs $\delta(f_2)$, $\delta(rm)$ and $\delta(r)$ are used to update the adaptable weights Q_1(i), Q_B(k), v_i, the value of the movement V_{rm}(j).

We assumed a discount factor $\gamma = 1$ and used the TD(λ) algorithm to update the weights at the end of each trial. The updating rule is the following:

$$Q_1(i) = Q_1(i) + \alpha \cdot \left(\lambda^{\Delta_{12}} \cdot \delta(f_2) + \lambda^{\Delta_{1rm}} \cdot \delta(rm) + \lambda^{\Delta_{1r}} \cdot \delta(r)\right) \cdot b_1(i) \tag{9}$$

$$Q_B(k) = Q_B(k) + \alpha \cdot \left(\lambda^{\Delta_{2rm}} \cdot \delta(rm) + \lambda^{\Delta_{2r}} \cdot \delta(r)\right) \cdot B(k) \tag{10}$$

$$V_{rm}(j) = V_{rm}(j) + \alpha \cdot \left(\lambda^{\Delta_{rmr}} \cdot \delta(r)\right) \tag{11}$$

In all the equations above α represents the learning rate. The value of Δ is the temporal interval between the relevant task events, i.e. the onset of f_1, the onset of f_2, the response movement and the reward. To mimic the task temporal structure we took $\Delta_{12} = 30, \Delta_{1rm} = 40, \Delta_{1r} = 45, \Delta_{2rm} = 10, \Delta_{2r} = 15, \Delta_{rmr} = 5$ (The temporal intervals are expressed in units of the time step, $dt = 0.1$ s). For the simulation, we use $\lambda = 0.95$ and $\alpha = 0.05$.

Correlation analysis

To estimate correlations between the RT and the phasic DA firing rate response to f_2 at the perceived difficulty of class k_0, we employed trials from all the presented classes k, weighting them according to their corresponding beliefs $B(k_0 \mid k)$. Given the data $\{x_i\}$ and $\{y_i\}$ (i=1, …, n), to calculate the weighted Pearson correlation coefficient between the data pairs (x_i, y_i) we defined the mean of $\{x_i\}$ (and of $\{y_i\}$) and the covariance between the pairs, weighted with the weight vector $\{w_i\}$, respectively as:

$$\langle x \rangle_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{12}$$

$$cov(x,y)_w = \frac{\sum_{i=1}^{n} w_i (x_i - \langle x \rangle_w)(y_i - \langle y \rangle_w)}{\sum_{i=1}^{n} w_i} \tag{13}$$

Then, the weighted Pearson correlation coefficient is given by:

$$corr(x,y)_w = \frac{cov(x,y)_w}{\sqrt{cov(x,x)_w cov(y,y)_w}} \tag{14}$$

For each class k_0, we measure correlations between the DA response during f_2 and the RT (Fig. 8A) as $corr\left(DA_{f2}, RT\right)_{B(k_0|k)}$. Trials from class one (k=1) were not included in the analysis because this class was distinctly easier for the animal, and it is not clear whether behavior and neural activity reflect motivation under such conditions.

Similarly, we define the weighted correlations between RT and phasic DA response during f_1, and between RT and the mean WM DA activity. In these cases, when the perceived frequency was f_1^0, we employed trials from all the presented f_1^k, weighting them according to their corresponding beliefs $b_1(f_1^0 \mid f_1^k)$. Equations (2-4) were used to compute $corr\left(DA_{f1}, RT\right)_{b(f_1^0|f_1^k)}$ (Fig. 8B). For this analysis we included all values of f_1. In all cases, we considered only correct trials.

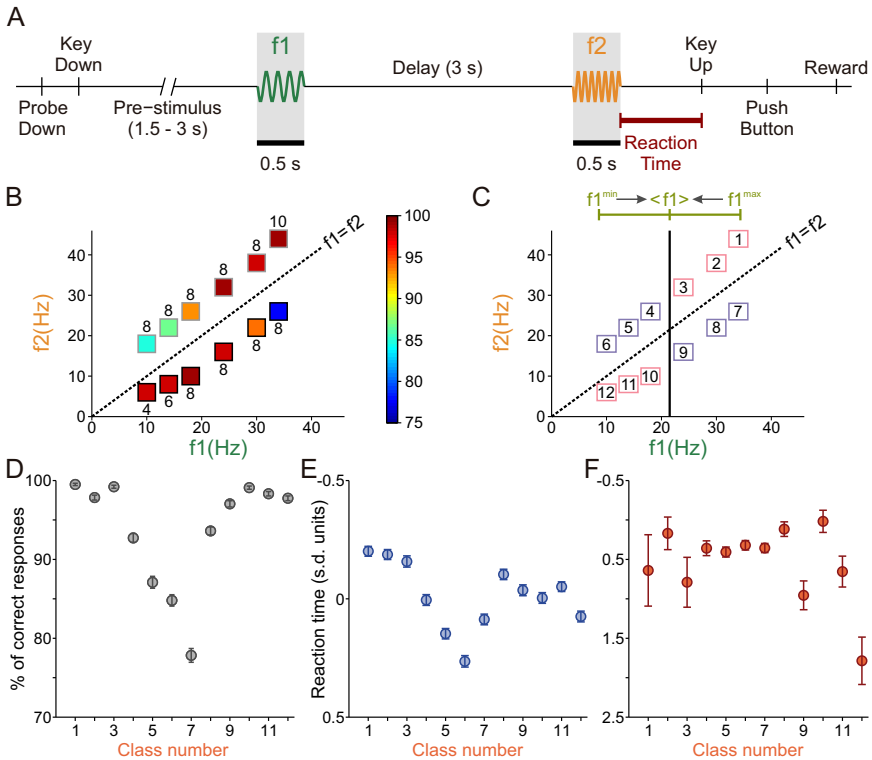

Figure 1. Discrimination task and contraction bias. (A) Schematic of the task design. The trial begins with the lowering of the mechanical stimulation probe (probe down, PD). The monkey reacted by placing its free hand on an immovable key (key down, KD). After a variable period (1.5-3.0 s) the probe oscillated for 0.5 s at the base (f_1) frequency. A 3-second delay after the first stimulus constitutes the working memory period (WM). The delay is followed by the presentation of the second stimulus (f_2, also the "comparison period"). The offset of f_2 signals the monkey to report its decision, and the time between offset and the monkey releasing the key (key up, KU) is known as the reaction time (RT, underlined in red). The monkey reports his decision with one of the two push buttons (PB) to indicate whether the comparison frequency was higher or lower than the base frequency. After that, the animal was rewarded for correct discriminations, or received a few seconds of delay for incorrect discriminations. (B) Stimulus set composed of frequency pairs (f_1, f_2) used in the task. Frequency values for f_1 are on the x-axis, while values for f_2 are on the y-axis. The diagonal dashed line represents stimulus equivalence (f_1=f_2). The box colors indicate percentage of correct trials, while the numbers above and below each box indicate the absolute magnitude difference (Δf) for each class. Red is for highest correct percentages; blue is for lowest. (C) The upper line (green) depicts the contraction bias effect, where the base stimulus is perceived as closer to the range center stimulus ($<f_1>$). The solid, vertical black line marks this center value. Numbers label classes, with the diagonal line marking stimulus equivalence (black, dashed). Upper diagonal ($f_1<f_2$) has bias benefit (pink squares) above range center, and bias obstruction (purple squares) below range center. Lower diagonal ($f_1>f_2$) has bias benefit below range center (pink), and bias obstruction above range center (purple). (D) The numbered classes (x-axis) from panel (C) by performance percentage (% of correct responses), representing the color values used in (B). (E) Reaction times (RT) in normalized units (standard deviation, s.d.) for correct trials per class. Longer average RT for lowest points, corresponding to bias obstructed classes. (F) RTs for incorrect trials per class. Classes with longest RTs corresponding to bias favored classes.

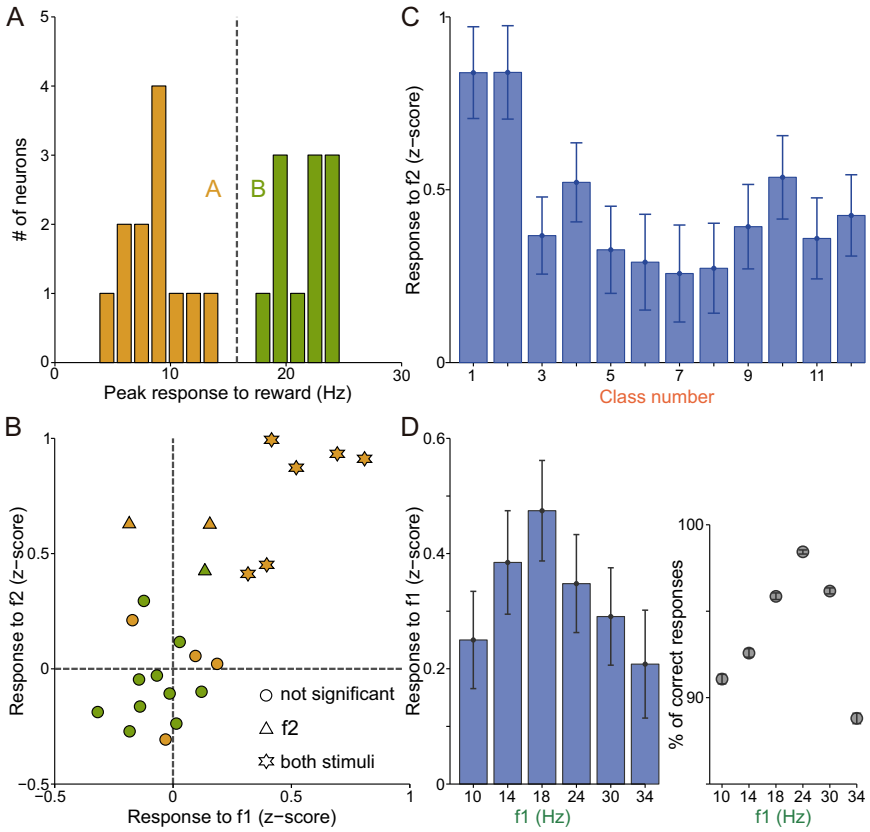

Figure 2. Neuron classification and phasic responses to both stimuli. (A) Histogram of the number of neurons as a function of peak reward responses in correct trials. Colored bars differentiate between 2 ranges of response peaks: orange bars for neurons in subpopulation *A* (low peak, n=12), green bars for neurons in subpopulation *B* (high peak, n=13). The dashed line (grey) between the two groups marks the criterion used (15 Hz). (B) Neuron subpopulations in a scatterplot. Each marker corresponds to one neuron. Colors mirror groupings from panel (A). Plotted by normalized activity for each stimulus (significance from Wilcoxon signed-rank test; Methods). Circles for neurons with no significant response. Triangles for neurons with significant responses for f_2. Stars for neurons with significant responses to both stimuli. (C) Normalized population responses to f_2, sorted by stimulus class (Methods). Error bars are ±1 SEM. (D) (Left) Normalized population response to f_1 plotted against f_1 frequencies. Error bars are ± 1 SEM. (Right) Percentage of correct responses for each f_1 frequency.

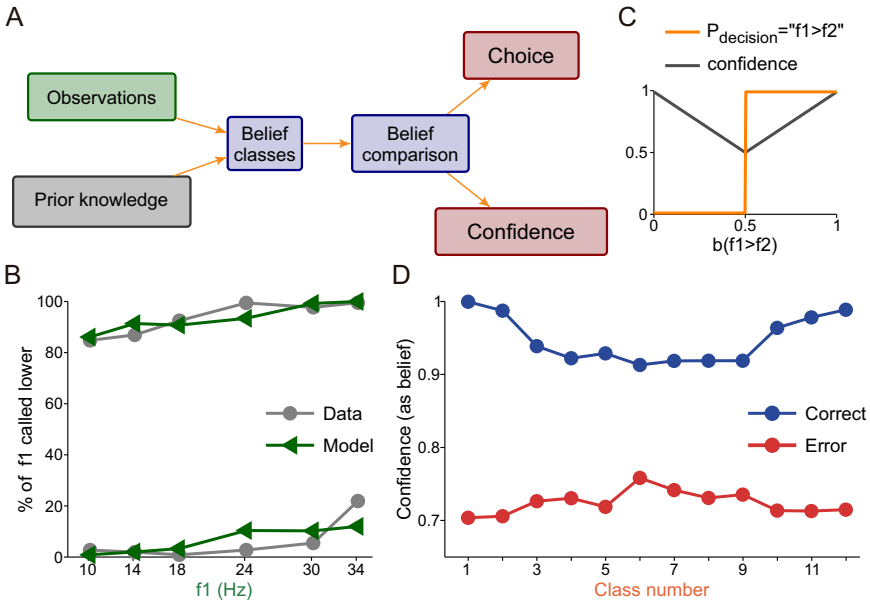

Figure 3. A Bayesian behavioral model and the relationship between the contraction bias and confidence. (A) Schematic of the Bayesian Behavioral Model: prior knowledge and observations are combined to obtain two comparison beliefs, b(H) ($f_1 > f_2$, Higher), and b(L) ($f_1 < f_2$, Lower). These beliefs are compared by magnitude in order to reach a decision. The decision can be decomposed into two components: Choice and Confidence. (B) Model fit of the percentage of times in which the base frequency (f_1) is called smaller. The behavioral data in grey (circles, lines), model predictions in green (triangles, lines). The line pairs for the 12 classes divided by the diagonal in Fig. 1D ($f_1 < f_2$, upper pair; $f_1 > f_2$, lower pair). Model best-fit parameter values were σ_1 = 5.5 Hz and σ_2 = 3.2 Hz. (C) Choices are made using *maximum a posteriori* (MAP) criterion (orange line). Confidence is defined as the maximum of the two beliefs b(H) and b(L) (grey line). (D) Bayesian model confidence predictions: correct trials in blue (circles, lines) and wrong trials in red (circles, lines).

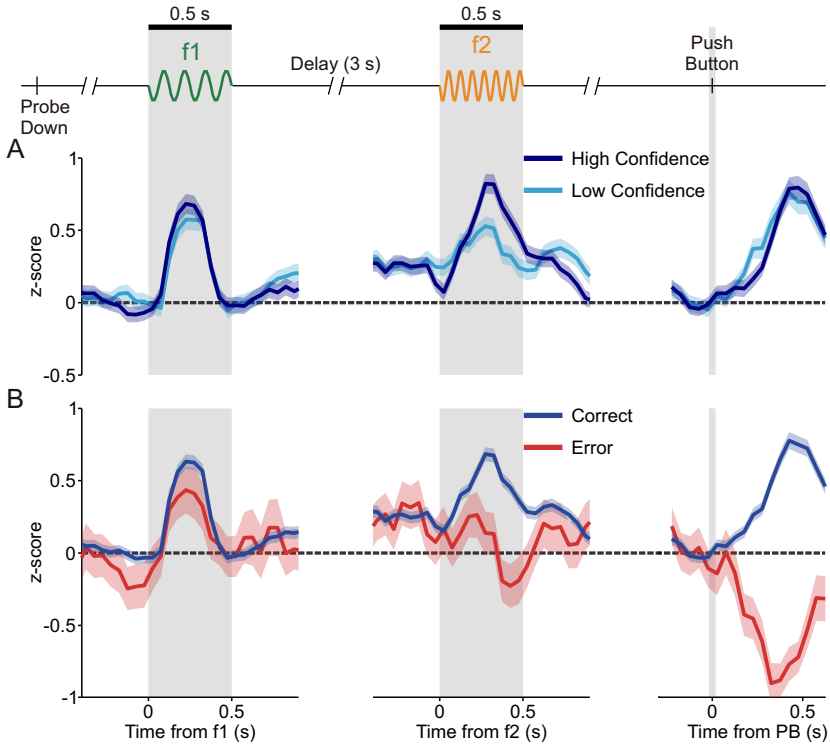

Figure 4. The phasic activation of DA neurons to the base stimulus, comparison stimulus, and reward delivery. Normalized DA population responses to different periods of the task (shadows ±1 SEM): (Left) Responses to f_1 (grey shaded region), aligned to f_1 onset. (Center) Responses to f_2 (grey shaded region), aligned to f_2 onset. (Right) Responses to reward delivery, after decision report (slender grey region), aligned to PB. (A) Responses for confidence groups: high (navy blue) and low (light blue) confidence. (B) Responses for trial outcome groups: correct (blue) and error (red) trials.

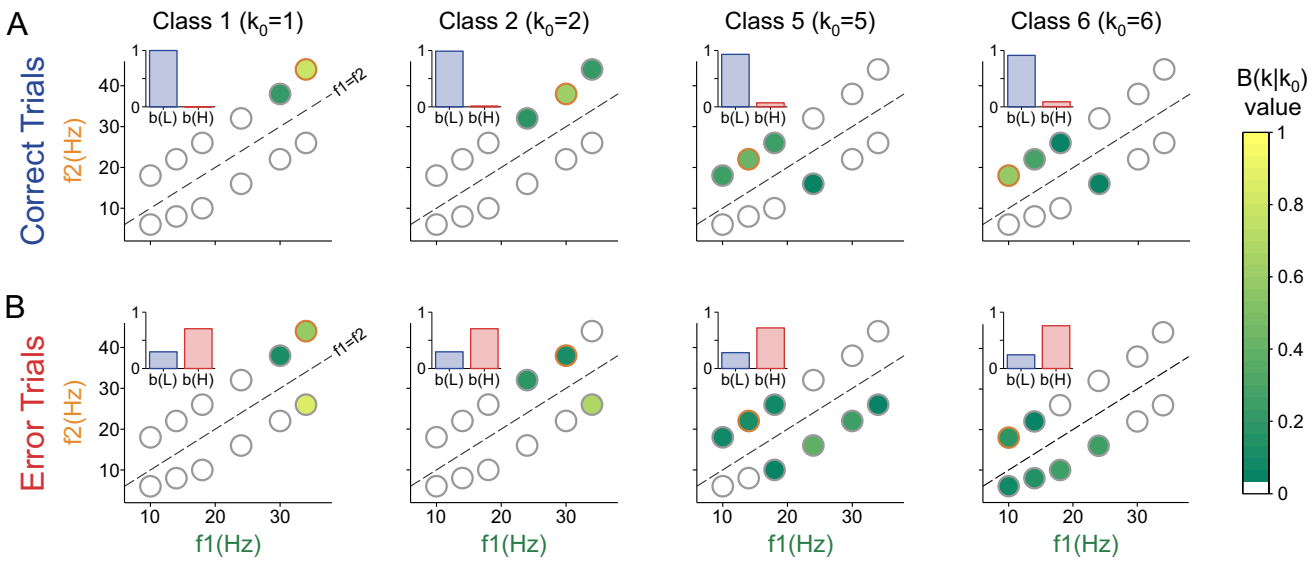

Figure 5. The Bayesian model relating confidence, DA activity and bias. Dispersion of the belief that class k was perceived instead of true class, k_0 ($B(k|k_0)$). Analysis results displayed for k_0={1,2,5,6}, all classes share the same correct response of ($f_1<f_2$). Favored classes are k_0={1,2}, disfavored classes are k_0={5,6}. Orange exterior circles highlight actual k_0 class, grey exterior circles highlight k class responses. All correct responses should be above the diagonal line. Plotted by f_1 (x-axis) vs. f_2 (y-axis), similar to Fig. 1B-C. Yellow for absolute belief ($B(k|k_0)$=1), green for intermediate belief ($B(k|k_0)$=0.5), white for no belief ($B(k|k_0)$=0). Inset graphs show the total belief values for b(L) ($f_1<f_2$, blue, correct) and b(H) ($f_1>f_2$, red, error). Confidence is the higher of these two beliefs (b(L) or b(H)). (A) Belief dispersion for correct trials. Note that b(L) (inset blue bar) is always greater. (B) Belief dispersion for error trials. Note that b(H) (inset red bar) is always greater.

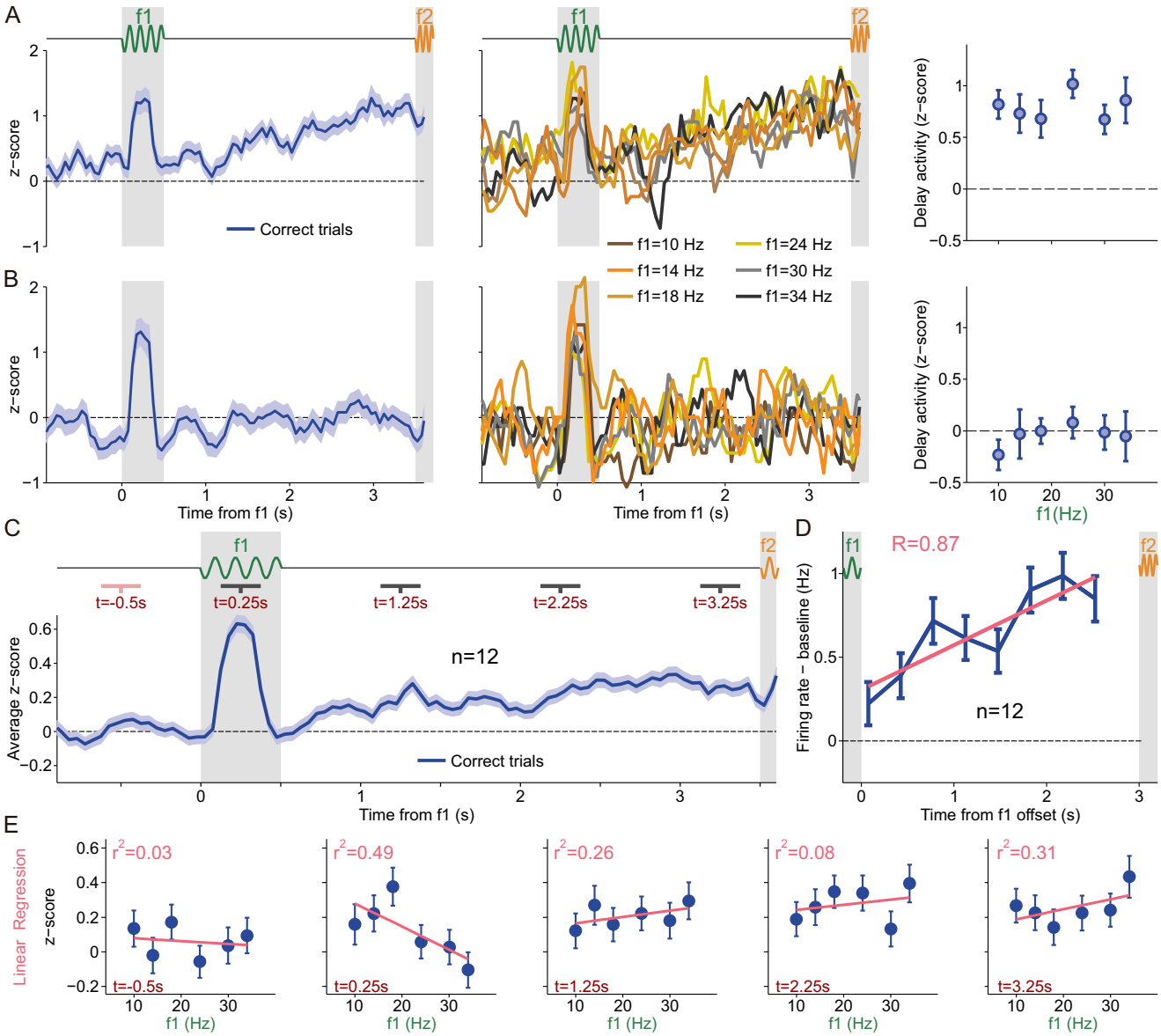

Figure 6. DA Activity during the delay period. (A-B) Normalized activity as a function of time for 2 example neurons, trials aligned to f_1 onset. Activity calculated in windows of 250 ms with steps of 50 ms (Methods). Neither example shows coding activity during WM. (Left) Activity from all correct trials was averaged (blue line, shadows ±1 SEM). Thick gray shaded area for f_1, thin grey shaded region for f_2. (Center) Same marked grey regions. Each curve indicates activity for one f_1 stimulus value. Curves with color closest to yellow for intermediate values, darkest curves for the extremum values. (Right) Average activity during the WM (delay) period as a function of f_1 values (error bars ±1 SEM). (A) Exemplary neuron with temporal dynamic during WM. (B) Single neuron without temporal dynamic. (C) Normalized activity as a function of time for subpopulation A between 1 second before f_1 onset through end of WM (blue line, shadows ±1 SEM). Trials aligned to f_1 onset. Grey regions mark same periods as for (A-B), activity calculated in same manner. Dark grey lines under task indicate analyzed periods after f_1 onset, pink line under task indicates the basal period before f_1 onset. Pink period compared against grey periods show significant modulations in each time segment (Wilcoxon signed-rank test; $p<0.001$, $p=0.002$, $p<0.001$ and $p<0.001$). (D) Mean population activity of DA neurons during the delay period, with correction for basal activity. Gray shadowed regions demonstrate stimulus presentation periods. The blue points for corrected firing rate in non-overlapping 350 ms periods (±1 SEM). Pink line displays linear regression across 8 discrete periods, where R is the determination coefficient. (E) Linear regressions (pink line) calculated in each of the 5 indicated regions from panel (C). Regressions over average activity for each f_1 stimulus (error bars for ±1 SEM). First graphic for the basal period before f_1 onset, proceeding in order of increasing time period from left to right. Times in bottom left corner (dark red), r^2 values are the determination coefficients and show the accuracy of each linear fit to the data. No linear regressions were found to have significant monotonic coding.

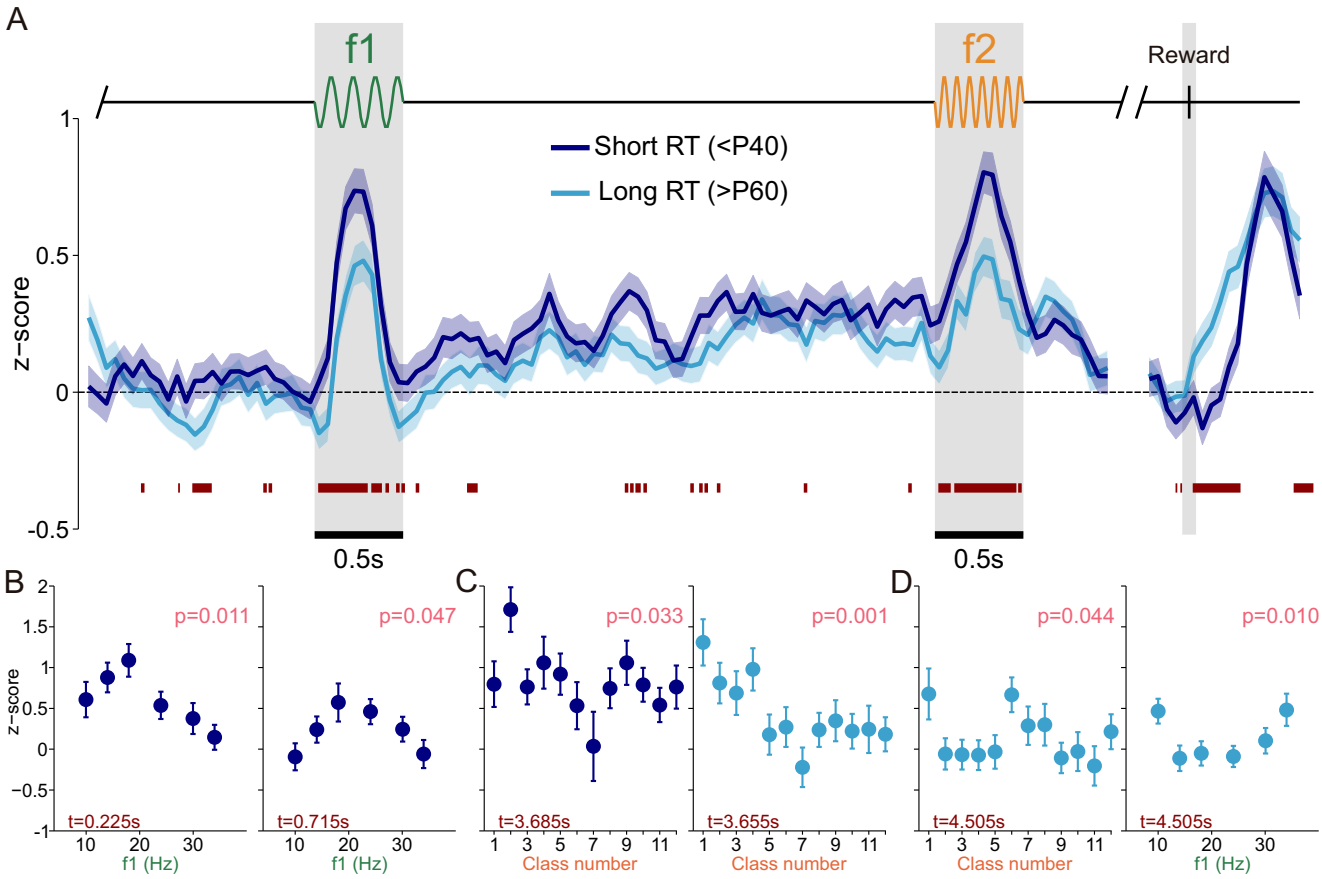

Figure 7. DA activity in short- and long-RT trials. (A) Normalized population activity as a function of time for correct trials, divided by reaction time duration: short-RT trials (RT<P40 (<40%), dark blue line, shadows ±1 SEM) and long-RT trials (RT>P60 (>60%), light blue line, shadows ±1 SEM). Red colored bar below curves marks significant differences between RT groups (AUROC, p<0.05; Methods). (B-D) Significance found with ANOVA test (Methods), marked by p value (pink), with time period in lower left corner (dark red). (B) Left: Population response (z-score) to f_1 onset ordered by f_1 frequency for short-RT trials (dark blue, ±1 SEM). Right: Population response after f_1 offset for short-RT trials (dark blue, ±1 SEM). (C) Population response at f_2 onset by class number for short- (dark blue, left) and long-RT (light blue, right) trials (±1 SEM). (D) Left: Population response after f_2 offset by class number for long-RT trials (light blue, ±1 SEM). Right: Population response after f_2 offset by f_1 frequency for long-RT trials (light blue, ±1 SEM).

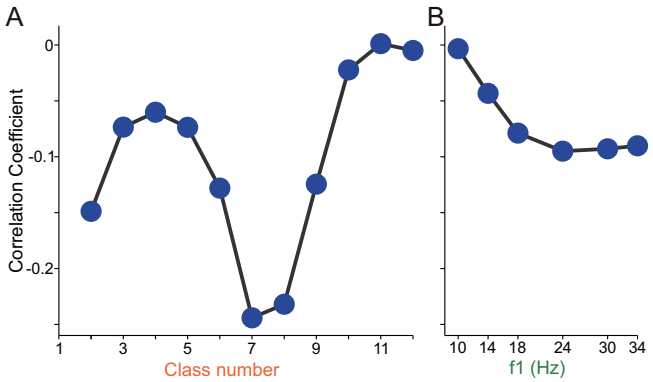

Figure 8. Dopamine reflects motivation on a trial basis. (A) Weighted Pearson correlation coefficient (PCC) between DA phasic response to class number and RTs. To estimate correlations, for each class k_0 we employed trials from all the classes k, each weighted with the corresponding belief, $B(k_0|k)$ (Methods). (B) Weighted PCC like panel (A) between f_1 and RTs. For each value of f_1 frequency a weighted PCC was calculated using the full data set, weighting each pair {DA, RT} with the corresponding belief $b_1(f_1^k)$ (Methods).

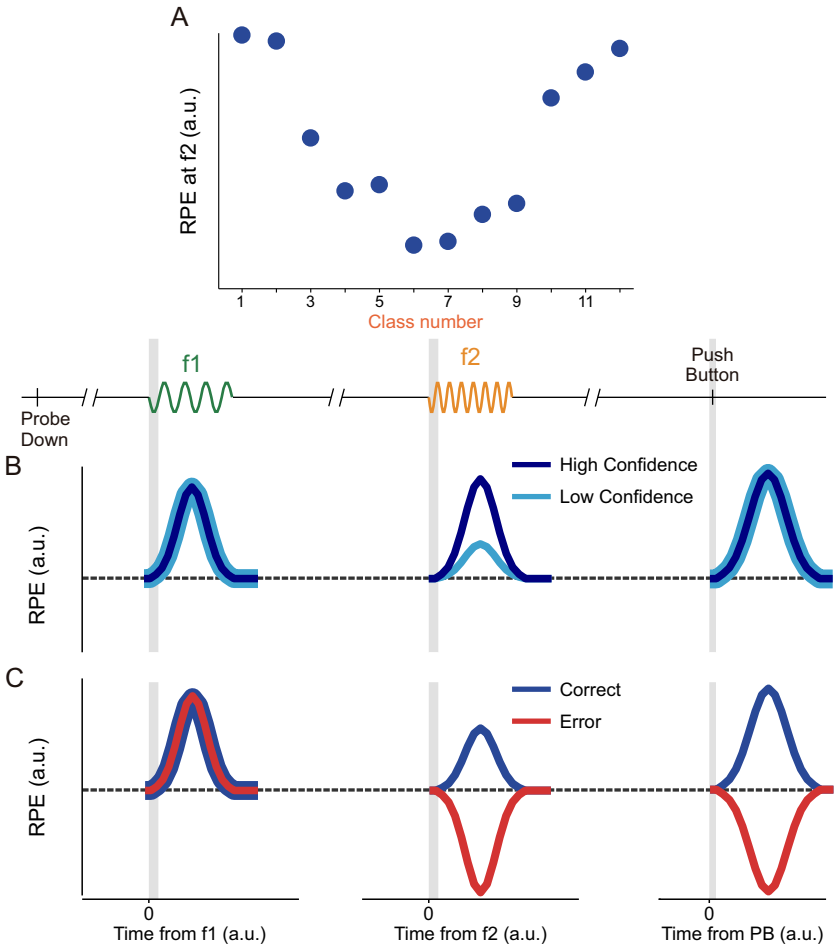

Figure S1. Reward prediction errors predicted by reinforcement learning model are similar to DA phasic responses. (Related to Figures 2 and 4). (A) RPE as a function of class number in correct trials, taken at the onset of f_2. A Gaussian filter was applied to the model predictions. Noise parameters σ_1=5.50 Hz for f_1 sampling and σ_2 = 3.2 Hz for f_2 sampling. (B) RPE signal calculated during 3 different periods with emulated data: f_1 presentation, f_2 presentation, and reward delivery after Push Button (PB, decision report). Dark blue curve represents the simulated high confidence group. Light blue curve represents the simulated low confidence group. Left: RPE after the onset of f_1 does not depend on the confidence level, observable in the overlap between both of our confidence group curves. Center: RPE after the onset of f_2 depends on choice confidence, favoring the high confidence group. The low confidence group (light blue) reaches less than half the max value observed for the high confidence group. Right: RPE after reward delivery peaks at a value independent of the confidence level, since the curves overlap perfectly. (C) RPE calculated during 3 different periods for simulated correct trials (blue curve) and simulated error trials (red curve). Left: RPE after the onset of f_1 is similar in correct and error trials, since the two curves overlap. Center: RPE after f_2 onset shows activation in correct trials and a strong depression in error trials (red curve). Right: RPE generated by the model after PB. Correct trials show an increase in RPE, while the error trials show a depression in RPE. The amount of change between error and correct trials is approximately equivalent.

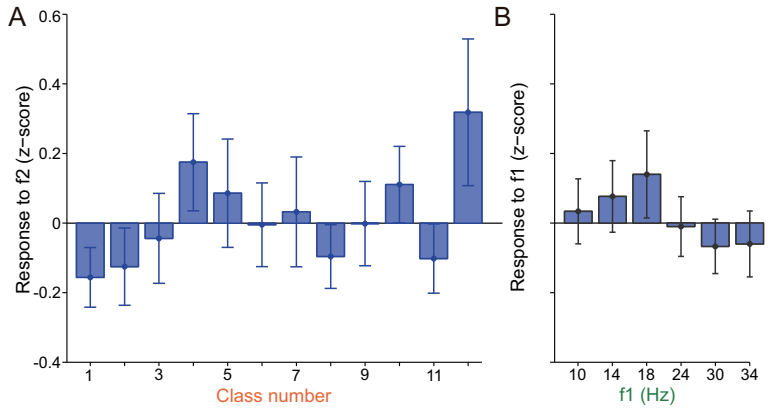

Figure S2. Phasic responses of neurons in subpopulation *B* (Related to Figure 2) (A) Population B response (z-score) to the onset of f_2 sorted by class number. Error bars are ± 1 SEM. There is no significantly normalized class dependence when f_2 stimulation begins. (one-way ANOVA, p<0.5). (B) Population B response (z-score) to the onset of f_1 sorted by f_1 frequency. Error bars are ± 1 SEM. Responses did not depend on f_1 (one-way ANOVA, p<0.5).

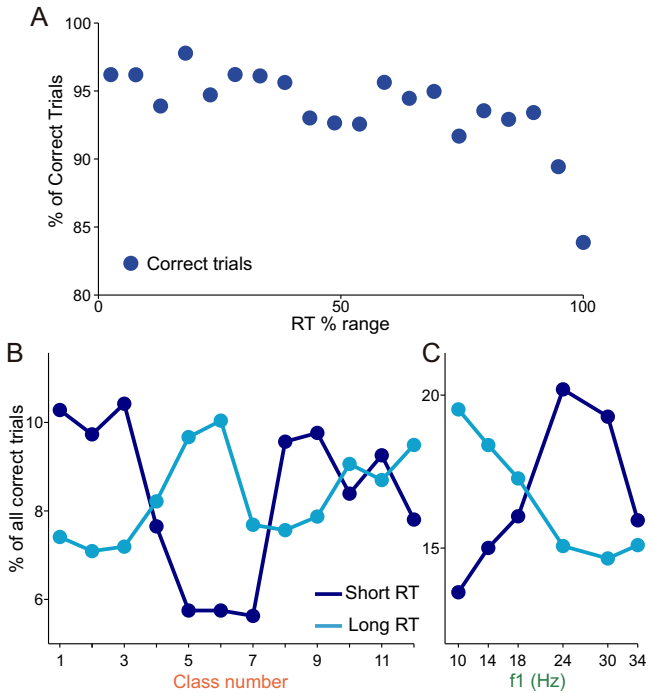

Figure S3. Correct trials based on RTs (A) The percentage of correct trials by the percentile of RT. As can be observed, longer RTs are associated with fewer correct trials. (B-C) The light blue color corresponds to the long-RT group, while the dark blue color corresponds to the short-RT group. (B) The percentage of correct trials by class number for each of the two RT groups. In short-RT trials easy classes appear more often than difficult ones. The opposite holds for long-RT trials. (C) The percentage of correct trials by the f_1 frequency intensity. In this case, short-RT trials have the most correct trials for f_1 frequencies which are perceived as easier, while in long-RT trials f_1 frequencies that are perceived as most difficult are more abundant.