Dopaminergic and frontal signals for decisions guided by sensory evidence and reward value

Armin Lak[1*], Michael Okun[2,3], Morgane Moss[1], Harsha Gurnani[1], Miles J Wells[1], Charu Bai Reddy[1], Kenneth D Harris[2], Matteo Carandini[1]

1 Institute of Ophthalmology, University College London, London WC1E 6BT, UK
2 Institute of Neurology, University College London, London WC1E 6BT, UK
3 Current Address: Centre for Systems Neuroscience, University of Leicester, Leicester LE1 7RH, UK
* arminlak@gmail.com

Abstract

Making a decision often requires combining uncertain sensory evidence with learned reward values. It is not known how the brain performs this combination, and learns from the outcome of the resulting decisions. We trained mice in a decision task that requires combining visual evidence with recent reward values. Mice combined these factors efficiently: their decisions were guided by past rewards when visual stimuli provided uncertain evidence, but not when they were highly visible. The sequence of decisions was well described by a model that learns the values of stimulus-action pairs and combines them with sensory evidence. The model estimates how sensory evidence and reward value determine two key internal variables: the expected value of each decision and the prediction errors. We found that the first variable is explicitly represented in the activity of neuronal populations in prelimbic frontal cortex (PL), which occurred during choice execution. The second variable was explicitly represented in the activity of dopamine neurons of ventral tegmental area (VTA), which occurred after stimulus presentation and after choice outcome. As predicted by the model, optogenetic manipulations of dopamine neurons altered future choices mainly when the sensory evidence was weak, establishing the causal role of these neurons in guiding choices informed by combinations of rewards and sensory evidence. These results provide a unified, quantitative framework for how the brain makes efficient choices when challenged with internal and environmental uncertainty.

Introduction

Making a decision often involves interpreting uncertain signals perceived by senses, and linking them with the rewards associated with the possible choices. For instance, when picking blackberries from a bush, one is confronted with uncertain visual signals and uncertain rewards. Their color varies from bright red -- clearly not ripe – to dark black -- clearly desirable. Many however are intermediate in color: the decision to pick them depends on one's taste, shaped by recent samples from the blackberry bush.

There has been substantial progress in understanding how the brain makes decisions based on sensory evidence, and on how it makes decisions based on past rewards. Using behavioral tasks that manipulated sensory evidence, it has been shown that the brain's sensory regions can interpret sensory signals to encode the probabilities of states of the world (Britten et al., 1992; Hernandez et al., 2000; Shadlen and Kiani, 2013). Neural correlates of the perceptual uncertainty associated with a choice have been established in multiple brain regions including parietal and frontal cortical regions (Hanks et al., 2015; Kepecs et al., 2008; Kiani and Shadlen, 2009; Middlebrooks and Sommer, 2012). In turn, behavioral tasks that manipulate reward parameters have revealed fundamental correlates of valuation and reward learning, with the relevant signals appearing in brain regions including frontal cortex and midbrain dopamine nuclei such as the ventral tegmental area (VTA) (Kennerley et al., 2009; Le Merre et al., 2018; Lee et al., 2012; Leon and Shadlen, 1999; Matsumoto et al., 2003; Morris et al., 2006; Padoa-Schioppa and Assad, 2006; Schultz et al., 1997).

It is unclear, however, how the brain combines representations of perceptual uncertainty and reward value, and uses these signals to make decisions and learn from the resulting outcomes. Efficient decision-making requires two fundamental computations: estimating the expected value of the choice, and revising those estimates using reward prediction error, i.e. the difference between prior expectation and the obtained outcome (Sutton and Barto, 1998). In the absence of perceptual uncertainty, expected values and prediction errors depend on past rewards alone. In contrast, when decisions require consideration of both sensory uncertainty and reward values, estimating the value of choice and prediction error requires combining immediately available, graded perceptual evidence with previously learned reward values (Dayan and Daw, 2008; Lak et al., 2017; Rao, 2010). Neuronal signals underlying such decisions should therefore reflect both perceptual evidence and reward values.

Signals for combining perceptual evidence and reward values may appear in VTA dopaminergic neuronal activity. Classical observations in putative dopamine neurons in VTA revealed responses elicited both by explicit cues that predict reward and by unpredicted rewards, resembling prediction errors defined in reinforcement learning models (Schultz et al., 1997). These observations were confirmed by subsequent studies that used genetic methods to examine cell type-specific dopamine neuronal responses (Cohen et al., 2012; Stauffer et al., 2016), and to establish the causal roles of dopamine neurons in guiding choices determined by past rewards (Hamid et al., 2016; Kim et al., 2012; Parker et al., 2016; Stauffer et al., 2016). There are also indications that putative dopamine neurons recorded in non-human primates reflect perceptual uncertainty (de Lafuente and Romo, 2011; Lak et al., 2017). It is not clear, however, if identified dopaminergic populations can integrate perceptual evidence and reward value. Critically, it is also not known whether dopamine responses can quantitatively act as a teaching signal to underlie both an animal's perceptual behavior and the influence of reward on it. Moreover, it is not known whether dopamine signals play a causal role during decisions guided by both reward values and sensory signals.

Signals for combining perceptual evidence and reward values may also appear in frontal cortex, for instance in the prelimbic region of rodent frontal cortex (PL). PL is one of the few cortical regions with functional projections to and from VTA dopamine neurons (Beier et al., 2015; Carr and Sesack, 2000; Morales and Margolis, 2017). Lesion or inactivation of PL renders animals insensitive to reward value and possibly impairs sensory detection (Killcross and Coutureau, 2003; Le Merre et al., 2018; Marquis et al., 2007). Some electrophysiological studies of PL activity in Pavlovian or operant conditioning suggested that PL responses are temporally related to sensory stimuli while indicating the presence or absence of future reward (Le Merre et al., 2018; Moorman and Aston-Jones, 2015; Otis et al., 2017). However, others suggested that PL responses are temporally related to actions, and are independent of trial-by-trial choice (Murakami et al., 2017). Thus, it is unknown whether the activity of PL neurons reflects the stimuli or the actions triggered by stimuli, and whether these neurons can combine perceptual evidence and reward values to signal the expected value of a decision.

A challenge with studying neuronal signals underlying decision behavior is that the relevant variables are internal: they are not directly observable, but rather need to be inferred from the choices. This inference requires quantitative models that uniquely define the relationship between these variables and behavior (Kepecs and Mainen, 2012; Shadlen and Kiani, 2013). In tasks guided by past rewards, such models have been successful in relating neuronal activity to moment-by-moment estimates of expected value or of prediction error (Bayer and Glimcher, 2005; Hamid et al., 2016; Parker et al., 2016; Samejima, 2005; Schultz et al., 1997). Quantitative models have also been developed for tasks guided by both past rewards and perceptual evidence (Lak et al., 2017; Rao, 2010). These models provide an opportunity to fit the observed choices and estimate the relevant internal variables quantitatively and moment-by-moment, so that they can be compared to neuronal activity.

We addressed these questions using tools that probe behavior, computations, and neuronal circuits. Inspired by tasks developed for humans and non-human primates (Nomoto et al., 2010; Rorie et al., 2010; Whiteley and Sahani, 2008), we designed a task for mice that required combining visual evidence and recent reward values, and showed that mice efficiently perform this combination to guide their trial-by-trial choices. We fit their choices with a simple model that combines current perception, memory of past rewards and learning from new rewards. The model provided moment-by-moment estimates of expected value and prediction error. Using electrophysiology, imaging and optogenetics, we found direct neuronal representations of these two variables: the first in populations of PL neurons and the second in VTA dopaminergic neurons. Further, we established that VTA dopamine signals are sufficient and necessary for guiding decisions informed by reward values and by perception.

Results

We begin by describing the behavioral task and the model that fits the observed choices. We then establish correlates for the model's internal variables in PL populations and in VTA dopamine neurons, and demonstrate the causal role of VTA dopamine activity.

Mice efficiently combine immediate perceptual evidence and learned rewards

To study decisions guided by sensory signals and learned reward values, we developed a behavioral task for head-fixed mice (Figure 1A-C). We first trained mice in a visual decision making task (Burgess et al., 2017) where in each trial, a grating appeared on the left or right side of the monitor, and the mouse indicated the stimulus position by steering a wheel with its forepaws (Figure 1A). Mice were rewarded with a water drop for correctly reporting the position of the stimulus (Figure 1B). To manipulate the perceptual difficulty of the task, we varied stimulus contrast across trials. After few weeks of training, mice could successfully perform the task, typically having near-perfect performance for easy (high-contrast) stimuli, while showing the expected graded relationship between performance and stimulus contrast (Burgess et al., 2017). We then manipulated reward values: in blocks of 50-350 trials, we rewarded successful choices on one side (left or right) with a larger water drop (1.2 vs 2.4 μl), and we switched this outcome contingency in consecutive blocks of trials (Figure 1C).

To maximize reward in this task, the mouse must combine sensory detection and estimation of available rewards. When the stimulus is unambiguous (at high contrast), the mouse should choose it regardless of its associated reward (choosing the other side would give no reward). Conversely, when the stimulus is hard to see (at low or zero contrast), the choices should be biased towards the side that would provide the larger reward. This strategy maximizes overall reward, as can be derived mathematically (Whiteley and Sahani, 2008), and with simulations (Figure S1A). Because reward value changed dynamically over blocks of trials, obtaining maximum reward requires the mouse to continuously learn from past choices and rewards.

Mice were successful in the task, efficiently combining perceptual evidence and past rewards into their trial-by-trial choices (Figure 1D-G). As expected from optimal performance (Whiteley and Sahani, 2008), the psychometric curves relating the proportion of rightward choices to stimulus contrast shifted sideways depending on relative reward value (Figure 1D,E). In other words, the value of expected rewards predominantly affected decisions in trials with contrasts near zero, while barely affecting decisions in easy, high-contrast trials (F = 25.5, $P < 10^{-10}$, 1-way ANOVA). Choice reaction times also depended on both stimulus contrast and reward size; increasing contrast or reward both decreased the reaction times (Figure S1B). The shift in psychometric curves developed gradually and continuously following changes in reward contingencies (Figure 1F,G). Following a change in reward contingency, mice took an average of 12 trials to shift their choices by 10% towards the side paired with larger reward (Figure 1G). These results demonstrate that mice can rapidly revise their inference of reward values from past trials and efficiently combine it with current perceptual evidence.

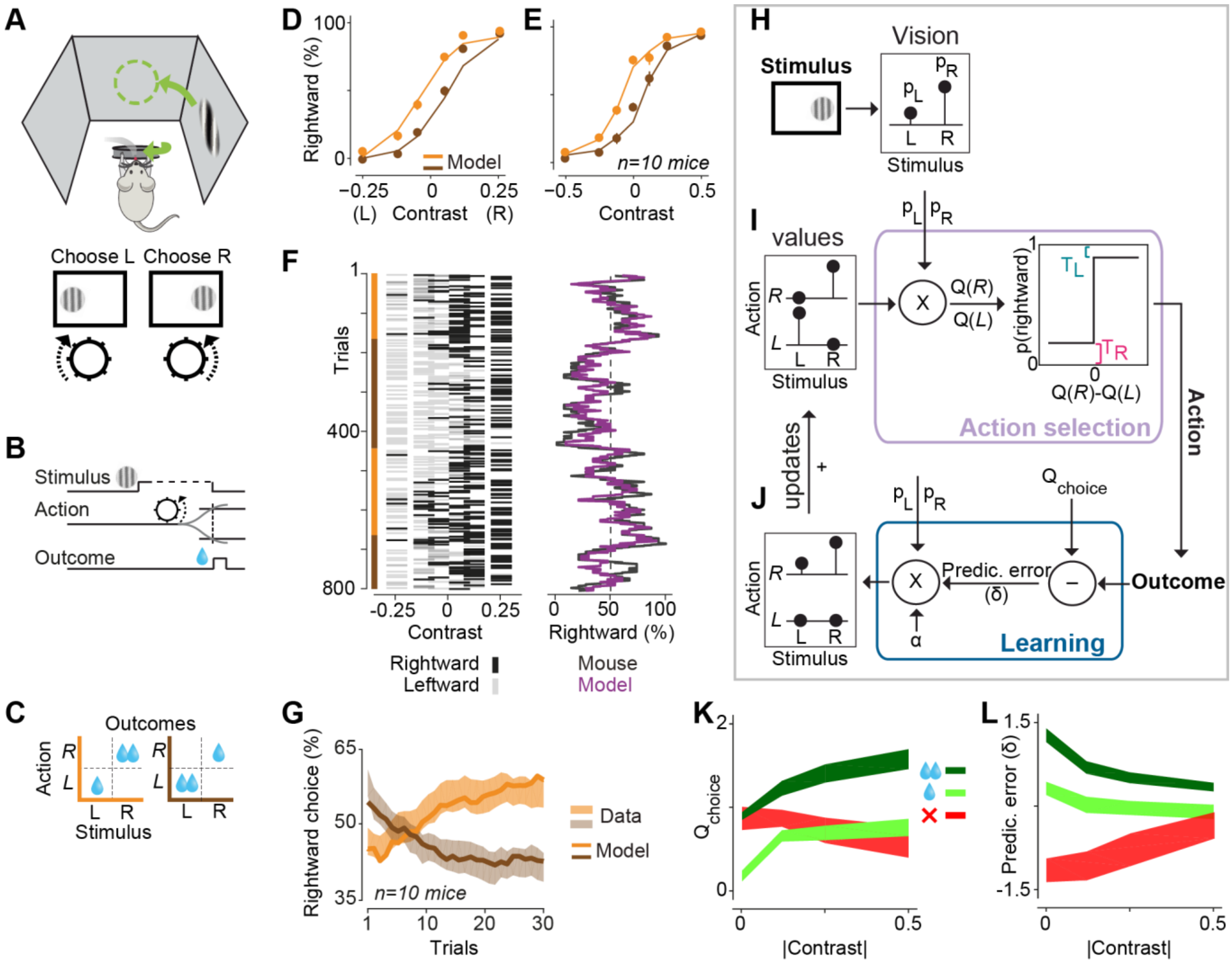

Figure 1. Behavioral and computational signatures of decisions guided by reward values and sensory evidence. A) 2-alternative visual decision making in mice. In each trial a mouse reports the position of a grating stimulus which appears on the left or right monitor by clockwise or counter clockwise steering of a wheel placed underneath their forepaws. B) Schematic of trial structure. C) Available rewards for successful choices were not equal and varied in consecutive blocks of trials. Within a block, correct choices to one side were rewarded with larger reward (2.4 vs 1.2 μl). Error trials were always followed by a 2s white noise. D) Behavior of an example animal in the task with unequal rewards. Solid curves are predictions of the model presented in H-J. Error bars are s.e.m across trials. E) Same as (D) but averaged across 10 mice for both observed choices and model fits. Error bars are s.e.m across animals. F) Left: Trial-by-trial choices in an example session. The bar on the right indicates the sequence of blocks. Right: Running average of trial-by-trial choices shown in left. Traces are shown for the mouse behavior (black) and for the model (purple). G) Learning curves of mice from the onset of blocks. Solid curves are predictions of the model shown in the next panels. H-J) A model for decision making based on immediate perceptual evidence and past rewards. H) In each trial the model computes sensory evidence i.e. the probability that the stimulus is in on the L or R side of the monitor. I) The model stores the value of stimulus-action pairs (the two-by-two matrix) and combines these with the sensory evidence to compute the values of taking L or R actions ($Q_L$and Q_R), and compares them to make a choice. J) The model computes a prediction error (δ) by comparing the trial outcome with the expected value of the choice (Q_{Choice}), and scales it by sensory evidence. Prediction error is then used to update learned values. K) Averaged estimates of Q_{Choice} of the model fits on choices of mice. Three trial types are shown: successful (correct) choices towards the large-reward side (dark green), successful trials towards the small-reward side (light green) and error trials towards the large-reward side (red). Width of shaded curves indicate s.e.m across animals. Trials with equal contrasts are grouped, regardless of whether they were presented on the left or right side (hence |contrast|). L) Similar to (K) but for reward prediction errors of the model.

A computational model that predicts trial-by-trial decisions

To capture the behavior of the mice and to make testable predictions about the underlying neural activity, we devised a model that captures how past rewards and current sensory evidence influence choices (Figure 1H-J, Figure S1C). The model is based on temporal difference reinforcement learning

(RL) and includes perceptual uncertainty and trial-by-trial learning of expected rewards from new outcomes (Lak et al., 2017). The model involves two key stages: action selection and learning.

The first stage of the model selects an action and computes its expected value Q_{Choice} (Figure 1H,I). This stage operates by combining two sets of quantities. The first set of quantities are the probabilities p_L and p_R that the stimulus is on the left or right side estimated by the visual system given the available but uncertain visual input (Figure 1H). The second set of quantities are the currently-stored expected rewards q_{sa} for each pair of stimulus *s* and action *a*. If the animal has learned the contingencies, the diagonal terms are small, and one of the off diagonal terms is larger than the other, Simple multiplication of these values provides $Q_L$and Q_R, the expected values of choosing left or right:

$$\begin{bmatrix} Q_R \\ Q_L \end{bmatrix} = \begin{bmatrix} q_{LR} & q_{RR} \\ q_{LL} & q_{RL} \end{bmatrix} \begin{bmatrix} p_L \\ p_R \end{bmatrix}$$

This stage then selects the action with highest expected value with probability $1 - T$, where T defines a (small) tendency to choose randomly (Figure 1I). This is known as a ε-greedy action selection rule (Sutton and Barto, 1998). The outcome of this stage is thus a choice (L or R) and the expected value of that choice,Q_{Choice}, defined as:

$$Q_{Choice}, = \begin{cases} Q_L \ if\ choice = L \\ Q_R\ if\ choice = R \end{cases}$$

The second stage of the model computes a reward prediction error δ to guide learning and thus inform upcoming decisions (Figure 1J). The reward prediction error is simply the difference between the obtained outcome and the expected value of the chosen option:

$$\delta = r - Q_{Choice}$$

After computing this value, this stage updates the stored q values weighted by the estimated probability of the stimulus (p_L and p_R) and a learning rate α (see Methods).

Note that, consistent with temporal difference RL framework, our model computes a prediction error also at the time of stimulus, because transiting from pre-stimulus state to the stimulus state changes the estimate of future reward (Figure S1C). This prediction error is proportional to Q_{Choice}, and is defined as the difference between Q_{Choice} and predicted reward at trial onset (a constant on average) (see Methods).

The model faithfully captured the behavior of the mice, offering a quantitative account of choices that require combining perceptual and reward information (Fig 1D-G). The model fitted the shift in psychometric curves seen after changing reward contingency (Figure 1D,E, *curves*), it accounted for trial-by-trial choices (Figure 1F, *purple trace*), and captured the time course of learning after a change in reward contingency (Figure 1G, *curves*). We examined the necessity of each model parameter using cross-validation, and found that the full model (with no dropped parameters) provided the best fit in 8 out of 10 animals (Figure S1D,E). We could thus use the model to estimate the internal variables that underlie choices: expected value of choice Q_{Choice} and prediction error δ.

The model makes quantitative predictions for how the expected value of choice Q_{Choice} should reflect reward size and uncertainty in obtaining the reward (Figure 1K). As might be expected, there is higher expectation to receive a reward when the stimulus has higher contrast. Indeed, for successful trials, Q_{Choice} is larger on the side that has the larger reward and grows as a function of stimulus contrast (Figure 1K, *dark green* and *light green*). Perhaps less intuitively, however, this dependence on contrast is reversed on error trials. When considering all trials in which the large-reward side was selected (which gives the most number of error trials), Q_{Choice} mildly decreases with contrast in error trials.

Indeed, errors made in the presence of high contrast stimuli reflect highly uncertain perceptual evidence (p_L is similar to p_R), or the model's very small tendency to choose randomly. In both these cases the estimated Q_{Choice} is small. A characteristic signature of Q_{Choice} is thus that it should increase with contrast in correct trials, but not in error trials.

Similarly, the model makes quantitative predictions for how reward size and uncertainty in obtaining the reward should affect the reward prediction error δ (Figure 1L). For successful trials, δ is larger upon receiving large reward and decreases with contrast, because predictions are more likely to be wrong when stimuli are hard to see (Figure 1L, *dark green* and *light green*). When considering all trials in which the large side were chosen, δ differs in the successful and error trials, in a manner opposite to Q_{Choice}: it grows with contrast in error trials (Figure 1L, *red*). This is yet another testable prediction of the model for how δ should behave during decisions guided by reward values and sensory evidence. We next set out to compare these quantitative predictions to the activity of frontal and dopaminergic neurons.

Prelimbic frontal neurons signal the expected value of the choice

To establish a neural correlate of expected value of choice Q_{Choice}— in its precise estimate provided by the model -- we recorded the activity of large populations in prelimbic frontal cortex (PL). We used high-density silicon probes to record from 1,566 neurons in the PL of 6 mice (Figure 2A). Approximately 20% (316) of these neurons responded to at least one task event (stimulus, action, outcome, Figure S2A; $P < 0.01$, signed rank test on responses prior and after each event).

The activity of responsive PL neurons tended to be tightly aligned to choice execution, i.e. wheel movement after the stimulus presentation (Figure 2B,C). Most neurons increased their firing before or around the time of choice execution (Figure 2C; 54%/24% of neurons showed significantly increased/decreased responses in the period -0.2 to 0.2 s around action onset, compared to baseline -0.4 to -0.2 s; $P < 0.01$, signed rank test). Consistent with a putative signal reflecting expected value, the responses tended to be the same for left choices and right choices (Figure 2C, only 5% of neurons encoded choice direction, $P < 0.01$, signed rank test).

Crucially, however, the firing rates of these neurons depended on stimulus contrast, reward size and perceptual accuracy (Figure 2D). When considering all the successful trials, the responses were higher in trials with higher stimulus contrast and in trials in which mice chose the large-reward side (Figure 2D, Figure S2B; contrast: Z = 5, $P = 10^{-6}$, reward size: Z = 2.6, $P = 0.009$, signed rank test). Moreover, when considering choices towards the large side (the condition that gives us the most error trials), neuronal responses were higher in successful trials than in error trials (Figure 2D, Figure S2B; Z = 5.8, $P = 10^{-8}$, signed rank test).

To confirm that the majority of PL neuronal responses occurred during choice execution, and to quantify these action-related responses, we set up a regression analysis (Figure S2C-H) (Park et al., 2014). We predicted the firing rate time course on each trial as a sum of three temporal kernel functions, time-locked to stimulus, action, and outcome, and scaled on each trial by coefficients to optimally fit the data (Figure S2C). The kernel for a particular task event represents the isolated neuronal response to that event with minimal influence from nearby events. The kernel waveforms were fit and cross-validated for each neuron, and subsequently, scaling coefficients that could change across kernels and across trials were fit (see Methods). Thus, the shape of the summed kernels captures variations in the neuronal activity that are due to the trial-by-trial differences in the temporal intervals between the events, and the coefficients reflect trial-by-trial variations in the neuronal response magnitude. The regression model accounted for a large fraction of neuronal variance (Figure S2D, upper panel). The action kernel was larger than the stimulus or outcome kernel in the majority of neurons, confirming that neurons were mainly driven by actions (Figure S2E; Z = 6.5, $P = 10^{-10}$, signed rank test on the maximum of kernel heights). Reduced regressions, which did not include the other type of task events, also revealed actions to be the most important event in driving PL responses (Figure S2D). Indeed, using a regression model that only included action events yielded predictions that strongly resembled the observed neuronal responses (Figure S2F-H). We could thus focus on the action-related kernel and use its trial-by-trial coefficients to summarize the trial-by-trial variations in the activity of PL neurons. Specifically, we could directly compare them to the expected value of choice Q_{Choice}.

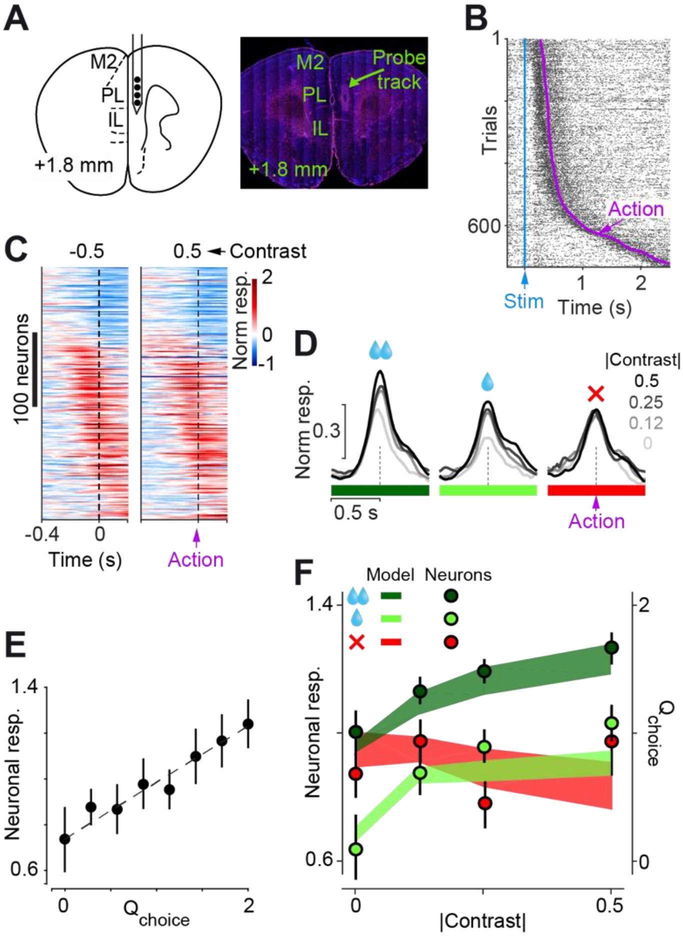

Figure 2. Prelimbic neurons reflect value of decisions requiring integration of perceptual evidence and reward values. A) Multichannel silicon probe recording from PL neurons (left) and example histological image showing the position of silicon probe in the PL frontal cortex (right). B) Raster plot showing responses of an example PL neuron aligned to the stimulus onset (blue line) and sorted according to choice reaction times (purple dots). C) Normalized (z-scored) activity of all task-responsive neurons aligned to the choice execution. i.e. onset of the wheel movement. The responses are sorted based on the latency of their maximum response. Left and right panels show responses for trials with high-contrast stimulus on the left and right side, respectively. D) Population neuronal activity during choice execution separated for stimulus contrast, the value of pending reward and successful/error choice. Left: responses in successful trials towards the large-reward side. Middle: responses in successful trials towards the small-reward side. Right: responses in error trials towards the large-reward side. E) Correlation of population neuronal activity (i.e. regression coefficients) and expected value of choice, Q_{Choice}, driven from the behavioral model. F) Averaged neuronal regression coefficients, separated based on stimulus contrast, the value of pending reward and perceptual accuracy (successful and error trials), overlaid on predictions of the model from Figure 1K. Error bars are s.e.m across neurons.

The trial-by-trial variations in action-related responses of PL neurons closely matched the model's estimates of Q_{Choice} , indicating that PL activity encodes the value of ongoing choices (Figure 2E,F). The trial-by-trial coefficients estimated from the neural responses showed remarkable correlation with Q_{Choice} estimated from the sequence of choices (Figure 2E, R^2=0.88, P = 10^{-4}, linear regression). Similar to Q_{Choice}, the neuronal coefficients increased with the size of the pending reward. Moreover, they increased with stimulus contrast, but only for correct choices (Figure 2F, *dark green* and *light green*). For incorrect choices, instead, they did not increase with contrast (Figure 2F, *red*). These observations match quantitatively the predictions of Q_{Choice} made by the model based on behaviour alone (Figure 1K). Further, control analyses showed that trial-by-trial coefficients are better correlated with the expected value of choices than with response vigor (Figure S2I; 45 vs 16 neurons, P < 0.01, linear regression). Taken together, these results indicate that the majority of PL neurons are responsive during choice execution, and that these neurons encode the value of ongoing choices, reflecting a precise combination of perceptual evidence and reward values.

VTA dopamine neurons combine perceptual evidence and reward history

To establish a neural correlate of prediction error δ – in its precise estimate provided by the model -- we measured the activity of dopamine neurons using fiber photometry. We expressed GCaMP6m in midbrain dopamine neurons of DAT-Cre mice (n=3) through viral vector injection, and implanted optical fibers above VTA to deliver excitation light and capture the fluorescence emitted by the dopaminergic neurons expressing GCaMP (Figure 3A). To ensure accurate measurement of Ca^{2+}

transients, we modified the task slightly to lengthen the trial duration: we trained mice to wait for an auditory go cue presented after the visual stimulus, before responding (Figure 3B).

As described previously, our model computes a prediction error not only at the time of outcome but also at the time of stimulus presentation, because the stimulus itself changes the estimate of future reward. This prediction error is proportional to Q_{Choice} and it is not causal for updating stimulus-action values (see Methods). At the time of outcome, the prediction error is the difference between the obtained reward and Q_{Choice}, which causally updates the stimulus-action values (Figure 1J). The model thus predicts that the signals that encode prediction error at the time of outcome will reflect the expected value of choice, Q_{Choice}, at the time of stimulus.

Consistent with these predictions, VTA dopamine cells were activated by both stimuli and outcomes (Figure 3C-F, Figure S3). Dopaminergic responses to stimuli did not reflect the position of the stimulus on the monitor (Figure 3C, $Z < 1.7$, $P > 0.08$ in 3/3 mice, signed rank test) but scaled with the size of pending reward (Figure 3D, $Z > 108$, $P < 0.004$ in 3/3 mice, signed rank test) and robustly reflected stimulus contrast (Figure 3E, left; $F > 209$, $P = 0$ in 3/3 mice, 1-way ANOVA). Actions did not significantly influence the neuronal responses (Figure S3, $Z < 1.4$, $P > 0.13$ in 3/3 mice, signed rank test). Dopamine responses to outcomes scaled with the size of reward (Figure 3F, right; $F > 300$, $P = 0$ in 3/3 mice, 1-way ANOVA).

Confirming these observations, neuronal regression analysis showed that dopamine responses depended strongly on stimulus and outcome, but were largely independent of the choice executions (Figure S3). To quantify the contribution of each task event in driving dopamine neuronal activity and examine trial-by-trial variations in neuronal activity, we again used a regression analysis (Figure S3A). We estimated a kernel for each task event (stimulus, action, outcome), together with a scaling coefficient for each kernel in each trial. A regression that only included stimulus and outcome events explained a substantial proportion of the variability in the neuronal responses (Figure S3B), and predictions of this regression showed close similarity to the observed neuronal responses (Figure S3C,D). These results indicate that VTA dopamine population neuronal activity is mainly driven by sensory stimuli and outcomes. We could therefore use the trial-by-trial coefficients of the stimulus and outcome kernels to estimate the size of dopamine responses at stimulus and outcome time, and compare them with Q_{Choice} and δ estimated from the model fitted to the behavior.

The dopamine neuronal responses at the stimulus time conformed with our model in a quantitative manner (Figure 3G,H). Trial-by-trial scaling coefficients at the stimulus time showed strong correlation with the expected value of decision, Q_{Choice} (Figure 3G; population: R^2=0.83, $P = 0.001$ and R^2>0.57, $P < 0.01$ in 3/3 mice, linear regression). Similar to Q_{Choice}, the neuronal coefficients increased with the size of the pending reward. Moreover, they increased with stimulus contrast, but only for correct choices (Figure 3H, *dark green* and *light green*). For incorrect choices, instead, they did not increase with contrast (Figure 3H, *red*). These observations match quantitatively the Q_{Choice} , and hence prediction errors at the stimulus time, estimated by the model.

The trial-by-trial variations in dopamine neuronal responses to outcomes closely matched the model's estimates of reward prediction errors (Figure 3I,J). The coefficients for the outcome time were strongly correlated with model-driven reward prediction errors (Figure 3I; population: R^2=0.97, $P = 10^{-6}$ and R^2>0.88, $P < 10^{-4}$ in 3/3 mice, linear regression). These coefficients depended on the stimulus contrast, reward size and perceptual accuracy, mimicking δ of the model (Figure 3J). Remarkably, at the time of outcome, despite the stimulus not being present on the monitor, the neuronal coefficients still reflected the stimulus contrast, in addition to reward size, and as expected, they differed in successful and error trials (Figure 3J). Thus, dopamine-specific neuronal activity resembles prediction errors of a RL model that combines perceptual uncertainty and reward values. Critically, model-driven prediction errors that could guide choices showed strong similarity to observed dopamine responses. These results indicate that midbrain dopamine neurons convey quantitative prediction errors that are suitable to act as teaching signals for guiding decisions requiring integration of perceptual evidence and reward values.

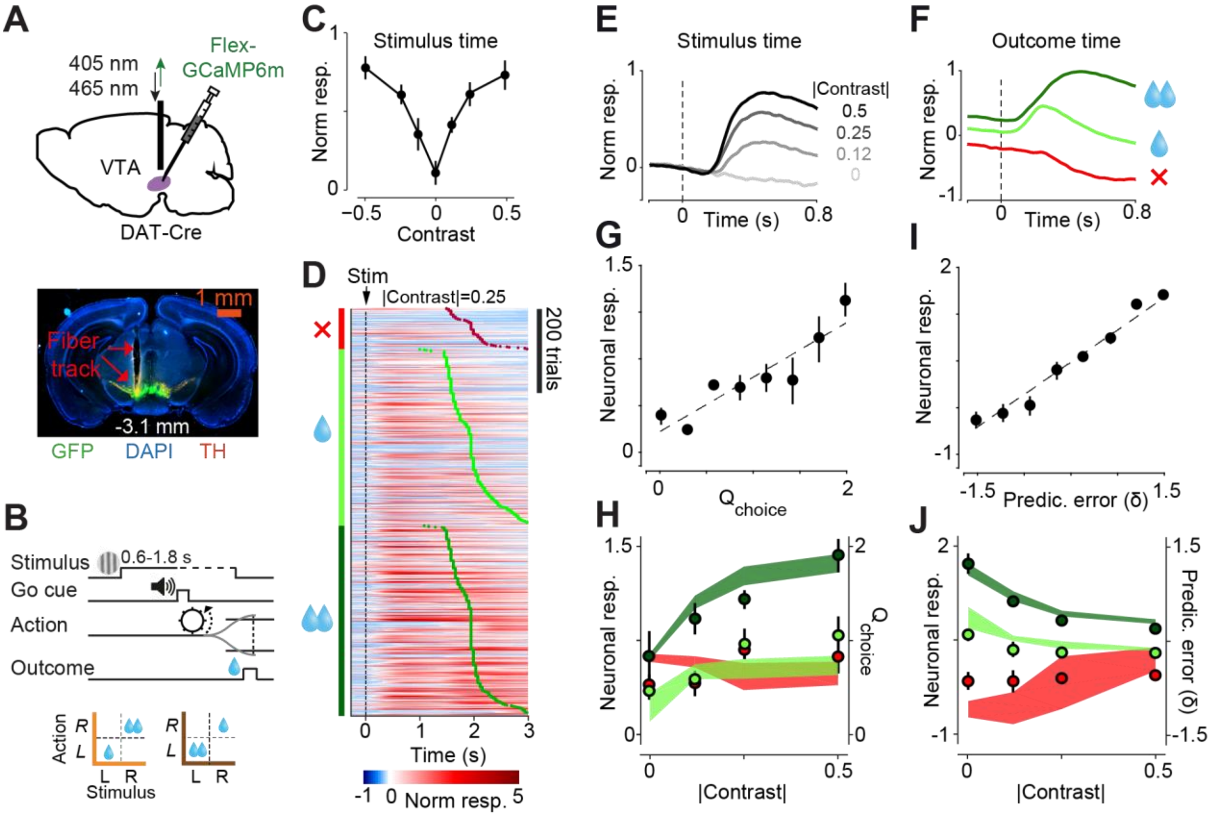

Figure 3. Dopamine prediction error responses integrate perceptual uncertainty and reward history. A) Upper: schematic of fiber photometry in VTA dopamine neurons. We injected viral vectors containing GCaMP6m to the midbrain of DAT-Cre mice and implanted a fiber over the VTA. Lower: Example histology slide showing expression of GCaMP in the midbrain and the position of implanted fiber above the VTA. B) The behavioral task. Similar to the task described previously but choice could be reported only after an auditory go cue. C) Average post-stimulus responses of rewarded trials for stimuli presented on the left or right side of the monitor. D) Trial-by-trial dopamine responses from an example animal in trials with |contrast| = 0.25. The responses are aligned to the stimulus onset (dashed line) and sorted based on the size of the pending outcome (shown on the left column) and the onset of outcome delivery (red, light green and dark green dots). E) Population dopamine responses (n=3 mice) aligned to the stimulus. F) Population dopamine responses aligned to the outcome onset. G) Correlation of estimated coefficients of neuronal regression at the stimulus time and model-driven estimates of expected value of choice, Q_{Choice}. H) Averaged regression coefficients at the stimulus time separated based on stimulus contrast, the value of pending reward and perceptual accuracy (successful and error trials), and overlaid on Q_{Choice} of the model (shaded curves). Error bars are s.e.m across recording sessions. I) Correlation of estimated coefficients of neuronal regression at the outcome time and reward prediction errors, δ, estimated from the behavioral model fitting. J) Same as (H) but for neuronal regression coefficients at the outcome time in relation to reward prediction errors, δ.

Causal roles of VTA dopamine neurons in choices based on perception and reward values

To establish the necessity and sufficiency of dopaminergic signals in choices based on perception and rewards, we used optogenetics (Figure 4A). We expressed Cre-dependent light-activated opsins in midbrain dopamine neurons of DAT-Cre mice. To activate dopamine neurons, we expressed Channelrhodopsin2 (Chr2, 9 mice) and to suppress dopamine neurons we expressed Archaerhodopsin3 (Arch3, 3 mice). We then implanted optical fibres above the left or right VTA (Figure 4A, Figure S4A). We trained the mice first in the regular task with equal water rewards on both sides, and when they reached stable performance, we added brief laser pulses to the water reward for correct choices toward one response side, and alternated the side in consecutive blocks of trials (Figure 4A).

Optogenetic activation and suppression had marked and opposite effects on the animal choices (Figure 4B, C). Optogenetic activation consistently shifted the psychometric functions towards the response side that was paired with laser pulses, both in individual animals and across the population (Figure 4B). This shift resembles the one we had observed with changes in water rewards (Figure 1D,E). Conversely, optogenetic suppression caused the opposite shift: curves shifted away from the response side paired with dopamine activation (Figure 4C).

The dopamine-dependent psychometric shifts developed over trials (Figure 4D), similar to the effects of manipulations of reward size, but in a faster manner; on average, after 8 trials mice shifted their choices by 10% towards the side paired with dopamine stimulation (Figure 4E). Also similar to the experiment with unequal water reward, choice reaction times were influenced by both stimulus contrast and the history of dopamine stimulation (Figure S4B). The psychometric shifts could be observed even in a trial-by-trial fashion in few experiments in which we paired water rewards with dopamine stimulation in a random subset of trials, rather than in a block of trials (Figure S4C).

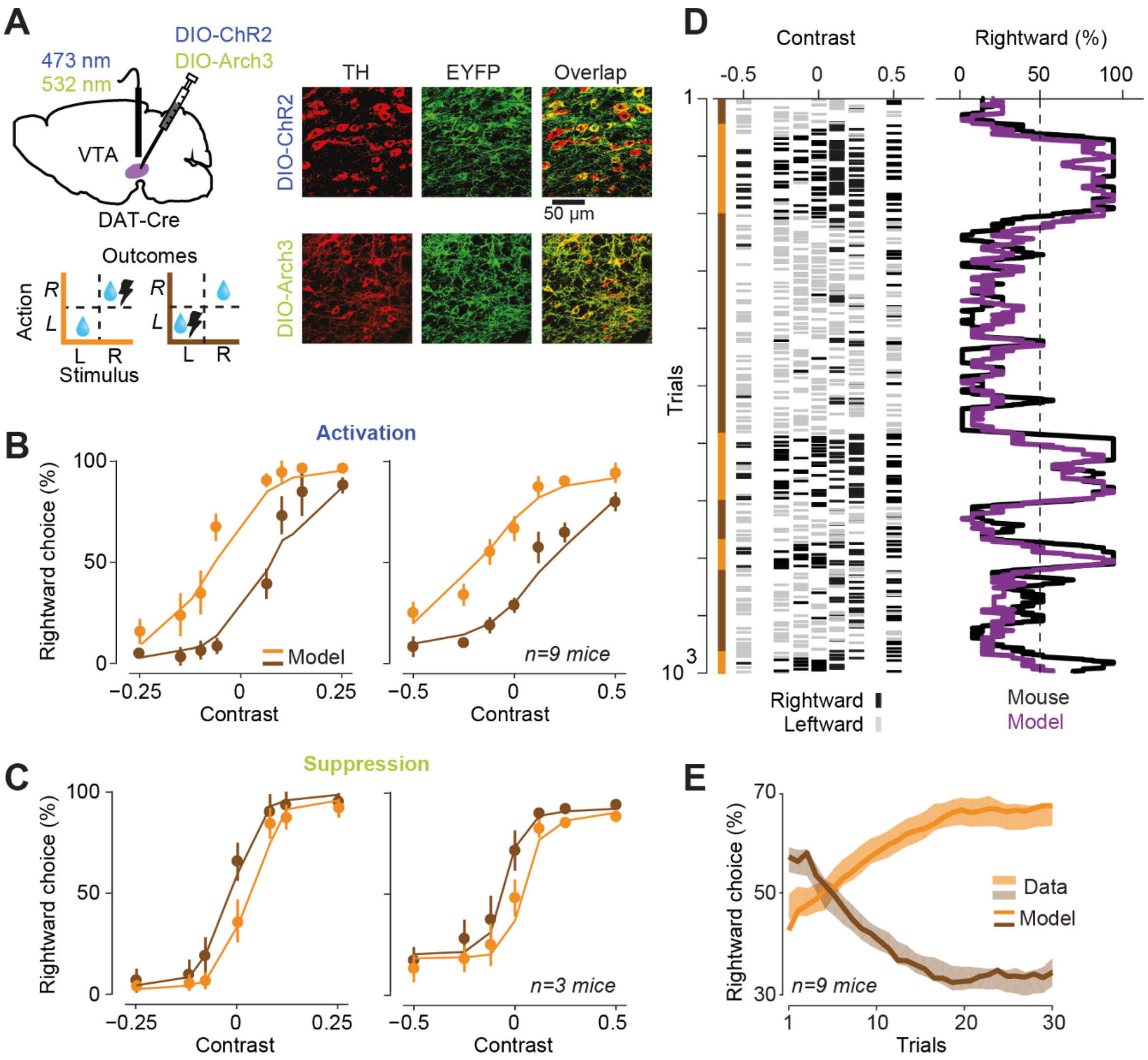

Figure 4. VTA dopamine neurons are necessary and sufficient for guiding choices informed by sensory evidence and rewards. A) We injected viral vectors containing ChR2 or Arch3 to the midbrain of DAT-Cre mice and implanted an optical fiber over the VTA. This resulted in expression of ChR2 or Arch3 in dopamine neurons (right panels). After reaching stable task learning using equal rewards, we started manipulating dopamine neuronal activity during the task. Similar to the experiments with unequal water, in consecutive blocks of trials, we paired correct choices toward one response side with brief laser pulses. The task timeline was similar to Figure 1B, i.e. it did not include a go cue. B) Psychometric curves were shifted towards choices paired with dopamine activation. The model accounts for dopamine-induced psychometric shifts (solid curves). Error bars are s.e.m across trials (left) or animals (right). C) Psychometric curves were shifted away from

choices paired with dopamine suppression. The model accounts for dopamine-induced psychometric shifts (solid curves). D) Left: Trial-by-trial choices in an example session. The bar on the right indicates the sequence of blocks. Right: Running average of trial-by-trial choices shown in left. Traces are shown for the mouse behavior (black) and for the model (purple). E) Learning curves of mice from the onset of blocks. Solid curves are predictions of the model.

These results reveal the necessity and sufficiency of dopamine neurons in guiding choices informed by both reward values and sensory perception. This conclusion is consistent with experiments that suggested that the activity of dopamine neurons reinforces previously rewarded actions (Hamid et al., 2016; Kim et al., 2012; Parker et al., 2016). However, the sideways shifts of the curves indicate that optogenetic manipulations were particularly effective when available sensory evidence was weak (activation: F = 3.8, *P* =0.01, suppression: F = 9.53, *P* =0.0006, 1-way ANOVA). Thus, the results demonstrate that dopamine neuronal activity influences subsequent choices mainly when perceptual uncertainty in the next trial is high.

Just as we have seen with changes in water rewards, the effects of dopamine manipulation were quantitatively predicted by our model (Figure 4B-E). Our model could account for both the psychometric curves and the trial-by-trial choices within blocks and at block transitions (Figure 4B-E, Figure S4C). Specifically, the model accounted for the effect of dopamine neuronal manipulation as a decrease or increase in the value of outcome for suppression and stimulation experiments, i.e. pronounced negative or positive prediction errors, respectively (Figure S4D,E). As the model indicates, these prediction errors update the stored value of stimulus-action pairs which can alter subsequent choice only if perceptual uncertainty is high, resulting in sideways shifts in psychometric curves.

We finally asked whether the activation of dopamine neuronal activity seen at the stimulus time is also causal to the task performance (Figure S4F-H). Both our imaging experiments and our model show that dopamine neurons are active not only at the time of outcome but also at the time of stimulus. The model requires the activity at time of outcome to be causal, so as to drive learning of associations between stimulus and action, but does not require that the activity at the time of stimulus to be causal. To test this, we optogenetically activated the dopamine neurons at the time of stimulus (rather than outcome). This manipulation decreased the animals' reaction times (Figure S4G), consistent with previous work showing a role for dopamine in motivating an animal to perform a task (Hamid et al., 2016). However, dopamine stimulation at the time of the stimulus did not influence the psychometric curves (Figure S4H, F = 0.09, *P* =0.99, 1-way ANOVA). Thus, while dopamine responses at the outcome time guide choices towards reward maximization in subsequent trials, dopamine responses to stimuli simply alters the animal's motivation to complete the trial.

Discussion

Taken together, our results provide quantitative evidence for the neural signals that support decisions that require combining uncertain sensory evidence with a history of reward. We trained mice in a task that requires this combination and found their choices to be well captured by a quantitative model which combines trial-by-trial perceptual evidence with stored values of stimulus-action pairs and uses uncertainty-dependent prediction errors to update the stored values. This model not only quantitatively describes mouse behaviour, but also predicts key internal variables which we were able to correlate accurately with neuronal activity in prelimbic cortex (PL) and ventral tegmental area (VTA). The majority of PL neurons fired at the time of choice execution and encoded the expected value of the chosen action. The activity of VTA dopamine neurons at the time of stimulus and outcome reflected not only reward size but also perceptual evidence, encoding prediction errors that scaled with both factors. Optogenetic manipulations of VTA dopamine neurons around the time of task outcome changed future choices in a manner consistent with the model's predictions.

As predicted by the model, the previously-learned reward values only affected choices at times when sensory evidence was weak. Similar behavior has been seen in humans and non-human primates tested in similar tasks (Nomoto et al., 2010; Rorie et al., 2010; Whiteley and Sahani, 2008). Indeed, when strong sensory evidence indicates that the side of small reward is correct, there is nothing to gain by selecting the other side, as no reward would come; but in zero-contrast trials, where the chance of either choice being rewarded is 50%, it is efficient to choose the side that is associated with the larger reward.

The model also successfully predicted an aspect in which the mouse behavior was not optimal. An observer that exploited the structure of our task would shift its psychometric curves after the very first time that reward was delivered but its size was unexpected, as this is certain to mark a block transition. Following a block transition, instead, both the mice and the model showed a gradual learning. This slow learning might reflect a “prior” that the rewards associated with different choices shift gradually, rather than suddenly. Indeed, slow fluctuations in reward values might be more common in nature than the abrupt shifts imposed by our experiment.

Our results show that prelimbic neurons were mostly responsive during choice execution, consistent with previous studies (Murakami et al., 2017), and their apparent post-stimulus responses are predominantly related to actions. These neuronal responses encoded the expected value of ongoing choice. PL activity was large for high-contrast trials, which is consistent with an encoding of expected reward-size since the probability the mouse chose correctly is largest for high-contrast stimuli. Nevertheless, contrast alone was insufficient to predict PL activity, as PL neurons responded more strongly for choices to the side paired with larger reward. Furthermore, when animals chose incorrectly despite a high-contrast stimulus, PL activity was low. In our model, incorrect choices for high-contrast stimuli occur either because the available perceptual evidence happens to be weak, or are due to model’s small tendency to make random choices. The expected value of the choice in both these conditions is low, consistent with the neuronal activity. Thus, PL activity at the time of choice execution reflects the value of upcoming reward, given the size of past rewards and the quality of perceptual evidence associated with the choice.

VTA dopamine neurons encoded the difference between the current and prior estimates of available rewards, taking into account the animal’s perceptual uncertainty. This activity was seen at the time of stimulus onset and reward delivery, but not at the time of movement onset. Dopaminergic activity at the time of stimulus was largest for choices towards large-reward high-contrast stimuli, and lowest for error trials, resembling previous observations in putative dopamine neurons in non-human primates (Lak et al., 2017). These responses are consistent with a model where the dopamine response to the visual stimulus also takes into account the mouse’s ongoing choice, indicating that dopamine responses to stimuli have rapid access to the quality of perceptual evidence used for choice computation. Dopamine activity at the time of outcome delivery reflected the difference between actual reward and the model’s expectation, taking into account perceptual uncertainty and reward history. The model used precisely this reward prediction error to update its internal value estimates. Consistent with the model, our optogenetic experiments supported a causal role for dopamine in learning the value of stimulus-action pairs. In particular, we found that optogenetic dopamine manipulation only affected choices in low-contrast trials, and this effect developed gradually over trials. In contrast, if dopamine was simply encouraging repeating of previously reinforced actions, then its reinforcing effect would not depend on perceptual uncertainty. Thus, dopamine neuronal activity are appropriate signals for teaching psychometric curves and for determining how reward value should influence perceptual choices. Interestingly, despite dopamine cells’ response to stimulus presentation, optogenetic dopamine stimulation at stimulus onset did not cause shifts in the animals’ choices but only decreased choice reaction times, suggesting temporally-dissociated roles of dopamine neurons in controlling learning and motivation.

The neuronal representation of perceptual uncertainty observed in PL and dopamine neurons are conceptually similar to previous reports of decision uncertainty in dorsal pulvinar (Komura et al., 2013), orbitofrontal (Kepecs et al., 2008; Lak et al., 2014) and parietal cortices (Kiani and Shadlen, 2009) suggesting that perceptual uncertainty signals spread through several brain regions, many of which have not been traditionally associated with perception. The observed neuronal representations in both PL and dopamine activity suggest that they might be functionally related. Given direct projections from anterior frontal cortex to VTA dopamine neurons (Beier et al., 2015), and recent evidence for causal roles of medial frontal cortex in shaping midbrain dopamine responses (Starkweather et al., 2018), our results suggest that perhaps dopamine neurons could receive an integrated reward predictive signal from PL, based on which they compute prediction errors appropriate for teaching a decision maker to harvest maximum reward in perceptually uncertain environments.

Neuroscience has at first studied the effects of perceptual uncertainty and reward learning in separate behavioral tasks: tasks probing perceptual decisions, where rewards are constant, and tasks probing the learning of past rewards, where perception is trivial. The resulting experiments have yielded data and models that are elegant yet segregated (Sugrue et al., 2005; Summerfield and Tsetsos, 2012). Our work brings these together and offers a behavioral and conceptual framework for circuit-level investigation of decisions guided by reward value and sensory evidence. Our findings reveal some of the subcortical and cortical computations that enable the brain to make efficient choices in inherently uncertain world where rewards are scarce.

Acknowledgements

This work was supported by Senior Investigator Awards from the Wellcome Trust (grant 205093) to M.C. and K.D.H., a Sir Henry Wellcome Fellowship (grant 106101) to A.L., and a Wellcome Trust PhD Studentship to H.G. We thank Rakesh Kotapati Raghupathy for histology, and other members of the laboratory for valuable comments. M.C. holds the GlaxoSmithKline/Fight for Sight Chair in Visual Neuroscience.

References

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron *47*, 129-141.

Beier, K.T., Steinberg, E.E., DeLoach, K.E., Xie, S., Miyamichi, K., Schwarz, L., Gao, X.J., Kremer, E.J., Malenka, R.C., and Luo, L. (2015). Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping. Cell *162*, 622-634.

Britten, K.H., Shadlen, M.N., Newsome, W.T., and Movshon, J.A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci *12*, 4745-4765.

Burgess, C.P., Lak, A., Steinmetz, N.A., Zatka-Haas, P., Bai Reddy, C., Jacobs, E.A.K., Linden, J.F., Paton, J.J., Ranson, A., Schroder, S.*, et al.* (2017). High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice. Cell Rep *20*, 2513-2524.

Carr, D.B., and Sesack, S.R. (2000). Projections from the rat prefrontal cortex to the ventral tegmental area: target specificity in the synaptic associations with mesoaccumbens and mesocortical neurons. J Neurosci *20*, 3864-3873.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85-88.

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

de Lafuente, V., and Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. Proceedings of the National Academy of Sciences *108*, 19767-19771.

Gunaydin, L.A., Grosenick, L., Finkelstein, J.C., Kauvar, I.V., Fenno, L.E., Adhikari, A., Lammel, S., Mirzabekov, J.J., Airan, R.D., Zalocusky, K.A.*, et al.* (2014). Natural neural projection dynamics underlying social behavior. Cell *157*, 1535-1551.

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.

Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C., and Brody, C.D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature *520*, 220-223.

Hernandez, A., Zainos, A., and Romo, R. (2000). Neuronal correlates of sensory discrimination in the somatosensory cortex. Proc Natl Acad Sci U S A *97*, 6191-6196.

Kennerley, S.W., Dahmubed, A.F., Lara, A.H., and Wallis, J.D. (2009). Neurons in the Frontal Lobe Encode the Value of Multiple Decision Variables. Journal of Cognitive Neuroscience *21*, 1162-1178.

Kepecs, A., and Mainen, Z.F. (2012). A computational framework for the study of confidence in humans and animals. Philosophical Transactions of the Royal Society B: Biological Sciences *367*, 1322-1337.
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 *324*, 759-764.
Killcross, S., and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb Cortex *13*, 400-408.
Kim, K.M., Baratta, M.V., Yang, A., Lee, D., Boyden, E.S., and Fiorillo, C.D. (2012). Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. Plos One *7*, e33612.
Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., and Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. Nat Neurosci *16*, 749-755.
Lak, A., Costa, G.M., Romberg, E., Mainen, Z.F., Koulakov, A., and Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. Neuron *84*, 190-201.
Lak, A., Nomoto, K., Keramati, M., Sakagami, M., and Kepecs, A. (2017). Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. Current biology : CB *27*, 821-832.
Le Merre, P., Esmaeili, V., Charriere, E., Galan, K., Salin, P.A., Petersen, C.C.H., and Crochet, S. (2018). Reward-Based Learning Drives Rapid Sensory Signals in Medial Prefrontal Cortex and Dorsal Hippocampus Necessary for Goal-Directed Behavior. Neuron *97*, 83-91 e85.
Lee, D., Seo, H., and Jung, M.W. (2012). Neural Basis of Reinforcement Learning and Decision Making. Annu Rev Neurosci *35*, 287-308.
Leon, M.I., and Shadlen, M.N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron *24*, 415-425.
Lerner, T.N., Shilyansky, C., Davidson, T.J., Evans, K.E., Beier, K.T., Zalocusky, K.A., Crow, A.K., Malenka, R.C., Luo, L., Tomer, R.*, et al.* (2015). Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. Cell *162*, 635-647.
Marquis, J.P., Killcross, S., and Haddon, J.E. (2007). Inactivation of the prelimbic, but not infralimbic, prefrontal cortex impairs the contextual control of response conflict in rats. Eur J Neurosci *25*, 559-566.
Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science *301*, 229-232.
Middlebrooks, P.G., and Sommer, M.A. (2012). Neuronal Correlates of Metacognition in Primate Frontal Cortex. Neuron *75*, 517-530.
Moorman, D.E., and Aston-Jones, G. (2015). Prefrontal neurons encode context-based response execution and inhibition in reward seeking and extinction. Proc Natl Acad Sci U S A *112*, 9472-9477.
Morales, M., and Margolis, E.B. (2017). Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. Nat Rev Neurosci *18*, 73-85.
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.
Murakami, M., Shteingart, H., Loewenstein, Y., and Mainen, Z.F. (2017). Distinct Sources of Deterministic and Stochastic Components of Action Timing Decisions in Rodent Frontal Cortex. Neuron *94*, 908-919 e907.
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.
Otis, J.M., Namboodiri, V.M., Matan, A.M., Voets, E.S., Mohorn, E.P., Kosyk, O., McHenry, J.A., Robinson, J.E., Resendez, S.L., Rossi, M.A.*, et al.* (2017). Prefrontal cortex output circuits guide reward seeking through divergent cue encoding. Nature *543*, 103-107.
Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature *441*, 223-226.
Park, I.M., Meister, M.L., Huk, A.C., and Pillow, J.W. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. Nat Neurosci *17*, 1395-1403.
Parker, N.F., Cameron, C.M., Taliaferro, J.P., Lee, J., Choi, J.Y., Davidson, T.J., Daw, N.D., and Witten, I.B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat Neurosci *19*, 845-854.
Rao, R.P. (2010). Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front Comput Neurosci *4*, 146.
Rorie, A.E., Gao, J., McClelland, J.L., and Newsome, W.T. (2010). Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey. PLoS One *5*, e9308.

Rossant, C., Kadir, S.N., Goodman, D.F.M., Schulman, J., Hunter, M.L.D., Saleem, A.B., Grosmark, A., Belluscio, M., Denfield, G.H., Ecker, A.S.*, et al.* (2016). Spike sorting for large, dense electrode arrays. Nat Neurosci *19*, 634-641.

Samejima, K. (2005). Representation of Action-Specific Reward Values in the Striatum. Science *310*, 1337-1340.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593-1599.

Shadlen, M.N., and Kiani, R. (2013). Decision making as a window on cognition. Neuron *80*, 791-806.

Starkweather, C.K., Gershman, S.J., and Uchida, N. (2018). The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. Neuron *98*, 616-629 e616.

Stauffer, W.R., Lak, A., Yang, A., Borel, M., Paulsen, O., Boyden, E.S., and Schultz, W. (2016). Dopamine Neuron-Specific Optogenetic Stimulation in Rhesus Macaques. Cell *166*, 1564-1571 e1566.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. Nat Rev Neurosci *6*, 363-375.

Summerfield, C., and Tsetsos, K. (2012). Building Bridges between Perceptual and Economic Decision-Making: Neural and Computational Mechanisms. Front Neurosci *6*, 70.

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

Tsai, H.C., Zhang, F., Adamantidis, A., Stuber, G.D., Bonci, A., de Lecea, L., and Deisseroth, K. (2009). Phasic Firing in Dopaminergic Neurons Is Sufficient for Behavioral Conditioning. Science *324*, 1080-1084.

Whiteley, L., and Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. J Vis *8*, 2 1-15.

Supplemental Figures

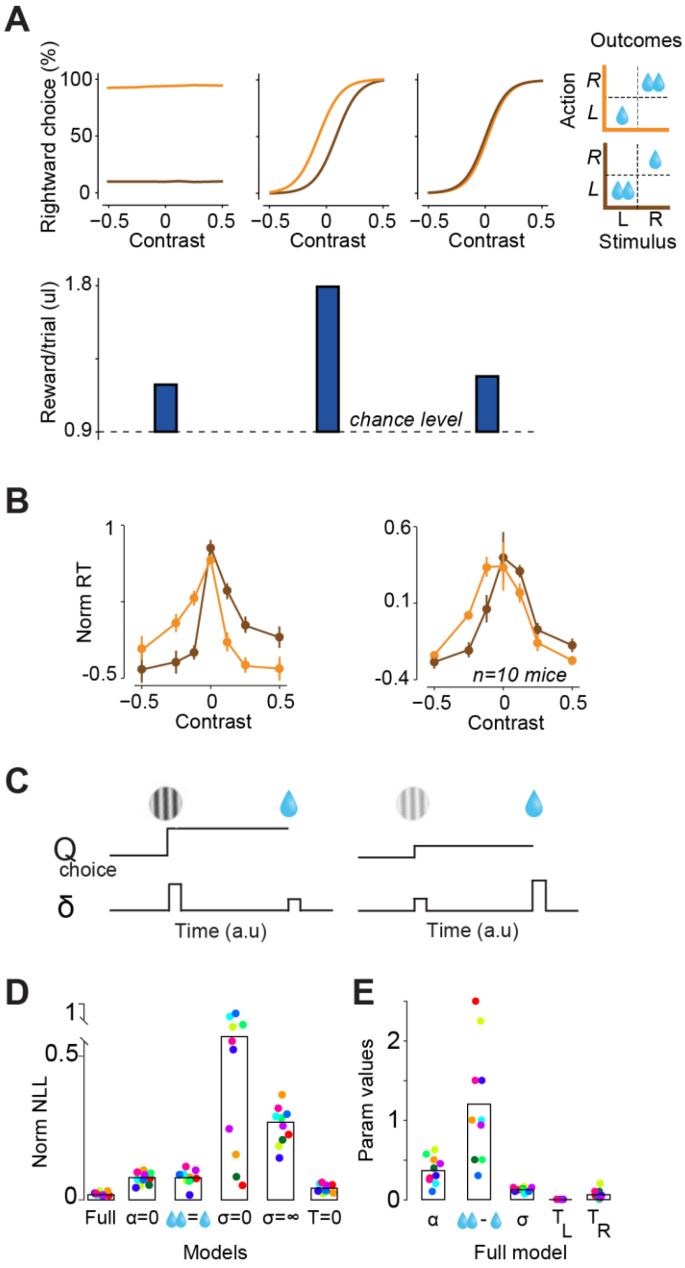

Figure S1. Mice efficiently combine reward values and perceptual evidence into their choices. A) Simulation of three agents that performed our behavioral task (top row) and their average reward harvest during 5000 trials. Left: an agent that made frequent choices toward the side paired with larger reward, regardless of the stimulus contrast. Middle: an agent that integrated past rewards and sensory evidence. Right: an agent that made choices according to the sensory stimulus, ignoring the reward size. Chance level in the lower panel indicates the reward harvest of an agent that made left and right choices in a random fashion. Simulations are performed using the model described in Figure 1. B) Choice reaction time of example mouse (left) and population of animals (right). For averaging reaction times across sessions and mice, they were z-scored. Reaction times reflected both stimulus contrast and reward size (contrast: F: 34, reward size: F=44.1, P < 10^{-10}, 2-way ANOVA,). Error bar are s.e.m across trials (left) or mice (right). C) Schematic of Q_{choice} and prediction errors of the model in a trial with high contrast and a trial with low contrast stimulus. D) Cross-validated negative log likelihood of various variants of the model. Full model contained all the parameters while each reduced model did not include one of the parameters. Cross-validation was 3 folds: sessions were divided to three, two of them were used for fitting and parameter estimation. The estimated parameters were then used to predict choices in sessions not used for parameter estimation. Circles with different colors represent different animals. F) Estimated parameters of the best model fitted on choices.

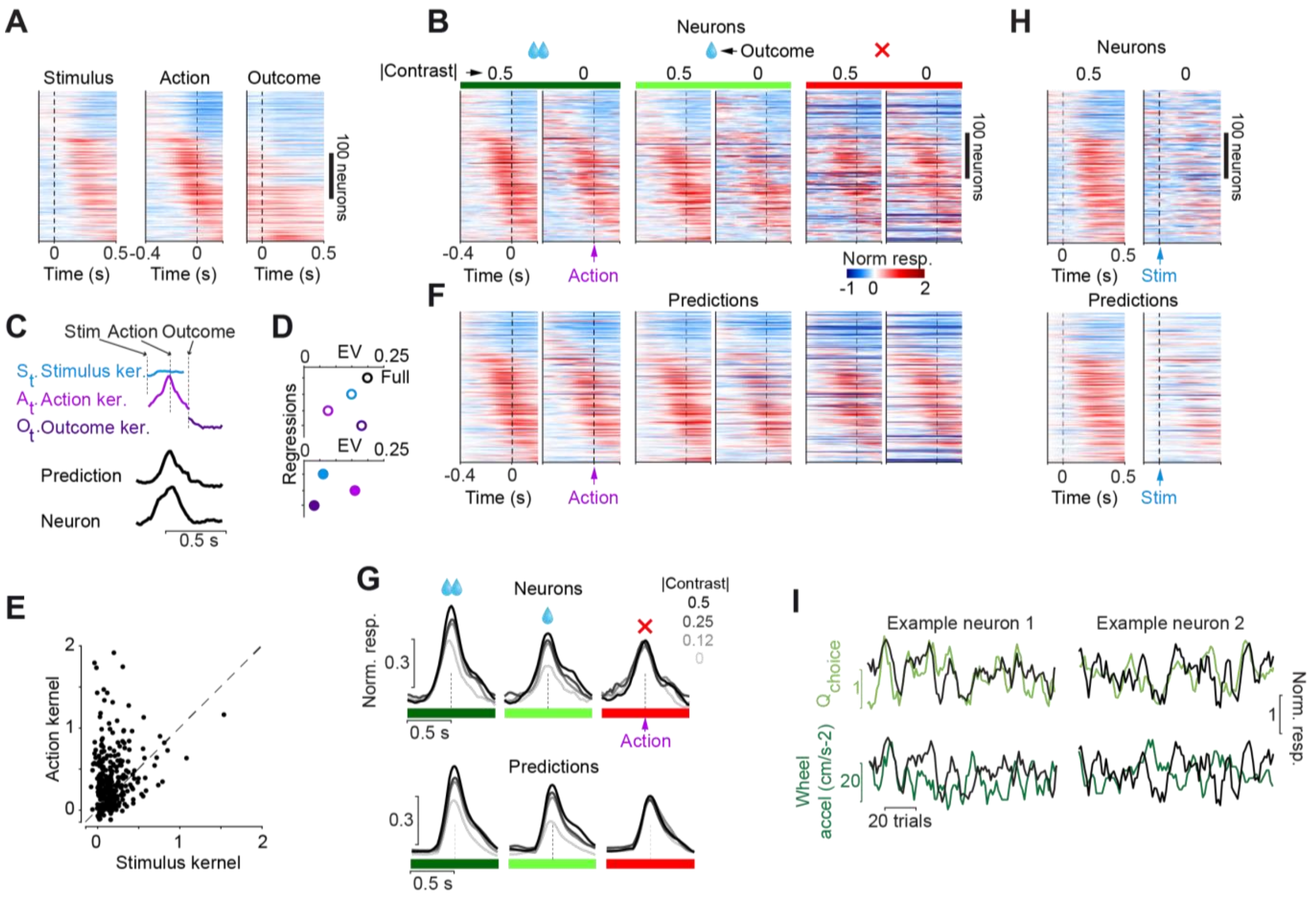

Figure S2. PL neuronal responses during the task. A) Normalized (z-scored) responses of all recorded task responsive neurons aligned to different task events and sorted according to the latency of the maximum responses in the middle panel. B) Normalized responses aligned to the action onset. Different panels show neuronal responses averaged across trials that included specific stimulus contrast and reward value, as shown above the panels. Dark horizontal blue lines in the panels of error trials (rightmost panels) reflect the fact that in few recording sessions animal performed very few error trials. C) Schematic of neuronal regression analysis. The regression estimates a temporal kernel for each task event, which are convolved with the events, scaled in each trial with a coefficient, and summed to produce regression predictions. D) Upper: Cross-validated explained variance (EV) averaged across neurons for the full regression and regression variants that did not include one of the kernels. Lower: EV of regressions that each included only one of the task events. Color code is as in (C). E) The size (maximum height) of action and stimulus kernels for the full regression. Each dot presents one neuron. Action kernels were often larger than stimulus kernels. F) Same as (B) but for predictions of the regression that only included action events. G) Population neuronal activity (upper panels) and predictions of the regression (lower panels) during choice execution separated for stimulus contrast, the value of pending reward and successful/error choice. Upper panels are identical to Figure 1D and are shown here for comparison with the predictions. H) Normalized neuronal responses aligned to the stimulus onset in all trials with |contrast| = 0.5 or 0. Neuronal responses are large after stimuli with high contrast but negligible when stimulus contrast = 0. These responses are captured by a regression that only included action events (lower panels). Thus, the response differences are due to different choice reaction times in high vs low contrast trials (see Figure S1B). Aligning neuronal responses to the choice onset reveals neuronal activity around the choice for both high and 0 contrast stimuli (for instance, leftmost panels in (B)). Therefore, the post-stimulus responses are predominantly due to actions that followed the stimuli. I) The fluctuation of responses of two example neurons vs estimated Q_{Choice} (upper plots) as well as wheel acceleration during choice, as a proxy of response vigor (lower plots). PL neurons were often better correlated with Q_{Choice} compared to wheel acceleration.

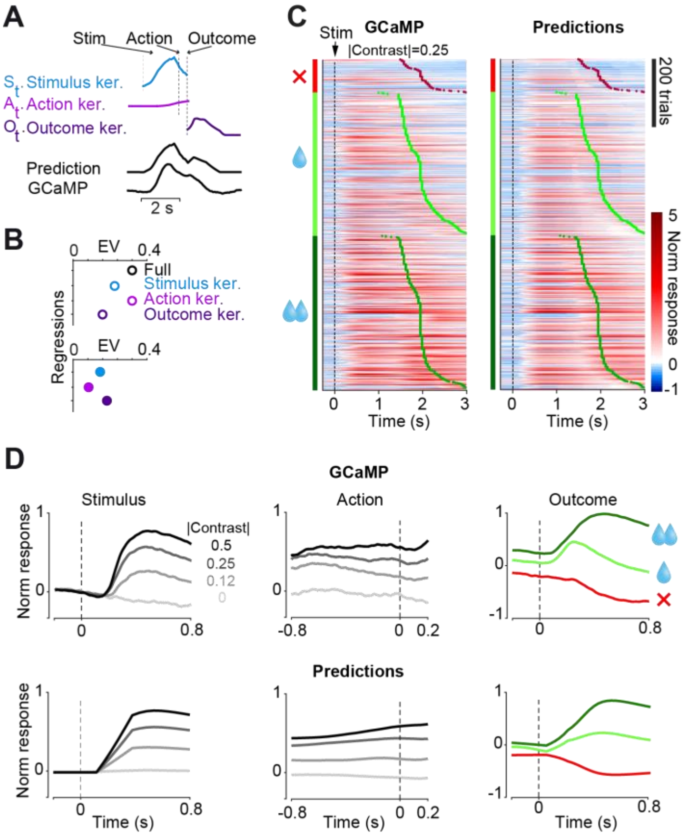

Figure S3. Dopaminergic neuronal activity during the task. A) Schematic of regression analysis. The regression estimates a temporal kernel for each task event, which are convolved with the event times, scaled in each trial with a coefficient and summed to produce regression predictions. B) Upper: Cross-validated explained variance (EV) averaged across mice for the full regression and regression variants that did not include one of the kernels. Lower: EV of regression variants that each included only one of the task events. Color codes are as in (A). C) Trial-by-trial dopamine responses (left) and predictions of the regression (left) in trials with |contrast| = 0.25. The responses are aligned to the stimulus onset (dashed line) and sorted based on the size of the pending outcome (shown on the left column) and the onset of outcome delivery (red, light green and dark green dots). The left panel is identical to Figure 3D, and is shown here for comparison with the predictions. D) Population dopamine activity aligned to the stimulus, action (i.e. first wheel movement after the visual stimulus onset) and outcome. Lower panels show the predictions of the neuronal regression.

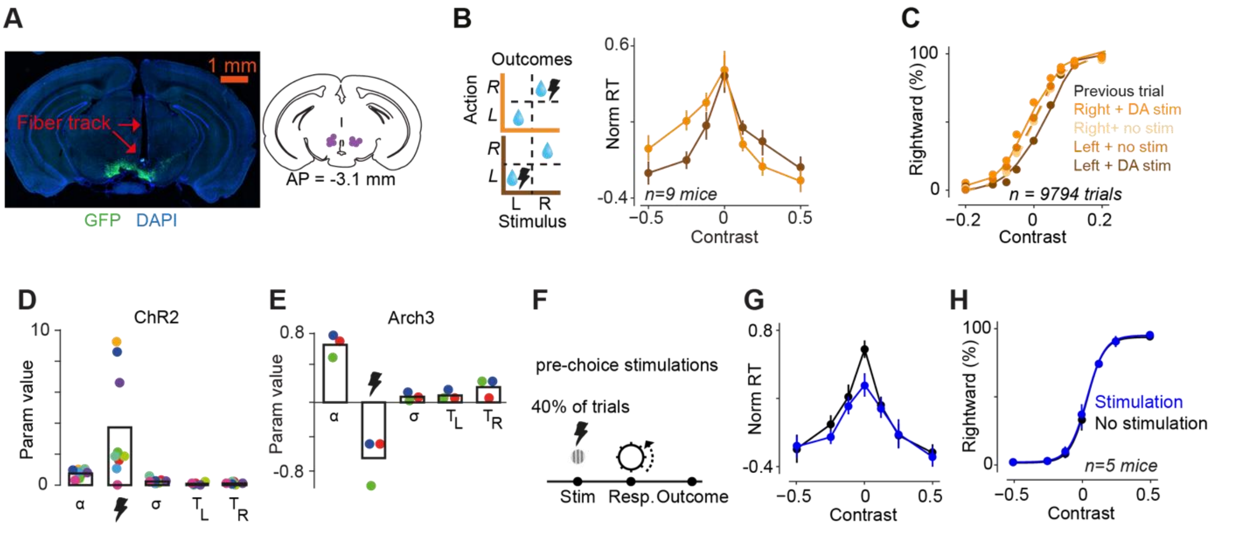

Figure S4. Optogenetic manipulation of VTA dopaminergic activity during the task. A) Left: Example confocal image showing placing of the implanted optical fiber above VTA as well as expression of ChR2 in midbrain of DAT-Cre mouse. Right: localization of implanted fiber tips from histological slices in difference mice. B) Choice reaction times separated for stimuli and dopamine stimulation blocks. C) Performance of an example mouse in experiments in which dopamine neurons were stimulated in randomly chosen successful trials (~30% of trials). Performance is plotted separately based on the chosen side and outcome of the previous trial. Stimulation of dopamine neurons in previous trial could shift the choices towards the side paired with such stimulation, only when the current perceptual evidence was weak. Solid curves are predictions of the behavioral model. D) Estimated parameters of the model fitted on experiments in which dopamine neurons were activated in consecutive blocks of trials. Circles with different colors represent different animals. E) Similar to (D) but for the experiments including suppression of dopamine neurons in consecutive blocks. F) Experiments for investigating the effect of stimulation of dopamine neurons prior to choice, i.e. during the stimulus presentation. Neurons were stimulated in 40% of randomly chosen trials. G) Choice reaction time in trials with and without pre-choice stimulation. H) Performance separated for trials with pre-choice stimulation and trials without stimulation.

Methods

Animals and surgeries

All experiments were conducted according to the UK Animals Scientific Procedures Act (1986). The data presented here has been collected from 25 mice (15 male) aged between 10-24 weeks. All mice were implanted with a custom-built metal head plate. To do so, the animals were anesthetized with isoflurane, and were kept on a feedback-controlled heating pad (ATC2000, World Precision Instruments, Inc.). Hair overlying the skull was shaved and the skin and the muscles over the central part of the skull were removed. The skull was thoroughly washed with saline, followed by cleaning with sterile cortex buffer. The head plate was attached to the bone posterior to bregma using dental cement (Super-Bond C&B; Sun Medical, Japan). For the electrophysiological experiments, we covered the exposed bone with Kwik-Cast (World Precision Instruments, Inc.), trained the animals in the behavioral task in the following weeks, and subsequently performed a craniotomy over the frontal cortex for lowering the silicon probes. For the fiber photometry as well as optogenetic experiments, after the head plate fixation, we made a craniotomy over the midbrain and injected viral constructs followed by implantation of the optical fiber, which was secured to the head plate and skull using dental cement. Post-operative pain was prevented by administering a non-steroidal anti-inflammatory agent (Rimadyl) on the three following days.

Behavioral tasks

Behavioral training started at least 7 days after the head plate implantation surgery. For mice which received viral injection, training started 2 weeks after the surgery. Animals were handled and acclimatized to head fixation for 3 days, and were then trained in a 2-alternative forced choice visual detection task, which we have previously described in details (Burgess et al., 2017). In each trial, a brief sound (0.1 s, 12 kHz) indicated that the mouse successfully kept the wheel still for at least 0.5 s and hence the trial has started. At the same time as the onset tone, the visual stimulus (a sinusoidal grating) appeared on either the left or right monitor and immediately afterwards the mouse could report its choice by steering the wheel located underneath its forepaws. Steering the wheel was translated to the movement of stimulus on the monitor (closed-loop condition): reward was delivered if the stimulus reached the center of the middle monitor (a successful trial) or a 2s white noise was played if the stimulus reached the center of the either left or right monitors (an error trial). As we previously reported, well-trained mice often reported their choices using fast stereotypical wheel movements (Burgess et al., 2017). In the initial days of the training, stimuli had contrast=1. Stimuli with lower contrast were introduced over days when the animal reached the performance of ~ 70%. After 2-3 weeks of training, the task typically included 7 levels of stimulus contrast (3 on the left, 3 on the right and zero contrast) which were presented in a random order across trials. We subsequently introduced unequal water rewards for correct choices: in consecutive blocks of 50-350 trials, correct choices to one side (left or right) were rewarded with larger reward (2.4 μl vs 1.2 μl of water) (Figure 1A-C).

Experiments involving optogenetic manipulation of VTA dopamine neurons had the same timeline as described above. However, instead of delivering extra water, correct choices to one side were paired with brief laser pulses in order to activate or suppress neuronal activity (Figure 4). In the experiments involving imaging of VTA dopamine activity, the task timeline slightly differed from above, allowing longer temporal separation of stimulus, action and outcome (Figure 3B). In these experiments, wheel movements immediately after the visual stimulus did not move the stimulus on the monitor and did not result in a choice (open-loop condition). Instead, an auditory go cue (0.1s) which was played 0.6-1.8 s after the stimulus onset started the closed-loop during which animals could report the choice. Wheel movements prior to go cue did not terminate the trial and we did not exclude these trials from our analysis.

The behavioral experiments were controlled by custom-made software written in Matlab (Mathworks) which is freely available at: https://github.com/cortex-lab/signals. Instructions for both the software as well as hardware assembly is freely accessible at: http://www.ucl.ac.uk/cortexlab/tools/wheel.

Electrophysiological experiments

We recorded the activity of prelimbic frontal cortex (PL) using multi-shank silicon probes in C57/BL6J mice. We paired chronic 32-channel silicon probe that had 2 shanks (Cambridge NeuroTech) with a moveable miniature Microdrive (Cambridge NeuroTech) and implanted it into the PL (n=6 mice). On the implantation day, we removed the Kwik-Cast cover from the skull, and drilled a small incision in the cranium over the frontal cortex, ML = 0.3 mm, AP = 1.8 mm (burr #19007–07, Fine Science Tools). The brain was protected with Ringer solution. We then lowered the probe through the intact dura using a manipulator (PatchStar, Scientifica, UK) to 1.4 mm from the dura surface. The final approach towards the target depth (the last 100–200 µm) was performed at a low speed (2–4 µm/sec), to minimize damage to the brain tissue. Once the probe was in its required position, we waited 10 minutes to let the brain recover from the mechanical strain of the insertion and then fixed the Microdrive on the head plate using dental cement. For reference signal we used a skull screw and silver wire which we implanted on the skull ~ 3-4 mm posterior to the recording site. At the end of each recording day we lowered the Microdrive 100 µm while monitoring the neuronal activity. Recordings were performed using the OpenEphys systems. Broadband activity was sampled at 30 kHz (band pass filtered between 1 Hz and 7.5 kHz by the amplifier), and stored for offline analysis. We implanted the animals after they fully learned to perform the task, performing the final stage of the behavioral task with performance above 70% for at least three sessions. Recorded spikes were sorted using the KlustaSuite software (Rossant et al., 2016) (www.cortexlab.net/tools), which involves three packages, one for each of three steps: spike detection and extraction, automatic spike clustering, and manual verification and refining of the clusters. Spike sorting was oblivious to task-related responses of the units.

Calcium imaging experiments

For measuring cell-type specific activity of dopamine neurons, we employed fiber photometry (Gunaydin et al., 2014; Lerner et al., 2015). We injected 0.5 µL of diluted viral construct GCaMP6m (AAV1.Syn.Flex.GCaMP6m.WPRE.SV40) into the VTA:SNc (ML:0.5 mm from midline, AP: -3 mm from bregma and DV:-4.4 mm from the dura) of DAT-Cre mice (B6.SJLSlc6a3tm1.1(cre)Bkmn/J, backcrossed with C57/BL6J mice). We implanted an optical fiber (400 µm. Doric Lenses Inc.) over the VTA, with the tip 0.05 mm above the injection site. We used a single chronically implanted optical fiber to deliver excitation light, and collect emitted fluorescence. We used multiple excitation wavelengths (465 and 405 nm) modulated at distinct carrier frequencies (214 and 530 Hz) to allow for ratiometric measurements. Light collection, filtering, and demodulation were as previously described (Lerner et al., 2015). For each behavioral session, least-squares linear fit was applied to the 405nm control signal, and the ΔF/F time series was then calculated as ((490nm signal – fitted 405nm signal) / fitted 405nm signal). Within-animal analyses were done by calculating *z*-scored ΔF/F.

Optogenetic experiments

DAT-Cre mice were used. We injected 0.5 µL of diluted viral constructs ChR2 (AAV5.EF1a.DIO.hChr2(H134R)-eYFP.WPRE) or Arch3 (rAAV5/EF1a-DIO-eArch3.0-eYFP) into the left or right VTA:SNc (ML:0.5 mm from midline, AP: -3 mm from bregma and DV:-4.4 mm from the dura) and implanted an optical fiber (200 µm, Doric Lenses Inc.) over the VTA, with its tip staying 0.5 mm above the injection site. We waited 2 weeks for virus expression and then started the behavioral training. After achieving stable task performance using equal water rewards, we introduced laser pulses with the following parameters: 473 nm and 532 nm for ChR2 and Arch3, respectively (Laserglow LTD), number of pulses: 12, each pulse lasting 10 ms and separated by 30 ms, laser power: ~10 mW (measured at the fiber tip). For the suppression experiment using Arch3, in few sessions we used a single 300 ms long pulse. In experiments involving manipulation of dopamine activity at the trial outcome, in consecutive blocks of 50-350 trials, correct choices to one side, L or R, were paired with laser pulses (Figure 4). In experiments involving trial-by-trial manipulations at the trial outcome, in 30% of randomly chosen correct trials, the choice was paired with laser pulses (Figure S4C). In both these experiments the laser was turned on simultaneously with the TTL signal that opened the water valve. For experiments involving stimulation of dopamine neurons at the stimulus time, in 40% of randomly chosen trials, we delivered laser pulses (with parameters described above) starting from either the onset of the stimulus or 200 ms before the stimulus onset (Figure S4F-H).

Histology and anatomical verifications

To quantify the efficiency and specificity of ChR2, Arch3 and GCaMP6M expression in dopamine neurons, we performed histological examination. Animals were deeply anesthetized and perfused, brains were post-fixed, and 60 µm coronal sections were collected. Sections were then immunostained with antibody to TH and secondary antibodies labelled with Alexa Fluor 594. For animals injected with ChR2 or Arch3 constructs, we also immunostained with antibody to eYFP and secondary antibodies labelled with Alexa Fluor 488 (Tsai et al., 2009) (Figure 3A, 4A, S4A). We confirmed viral infection efficiency and specificity in 135 confocal images collected from 14 (out of 15) mice injected with ChR2, Arch3 or GCaMP6M.

To determine the anatomical location of the implanted fibers, the tip of the longest fiber track found was used, which is represented in the corresponding Paxinos atlas slide (Figure 3A, S4A). For verifying the position of silicon probe, the coronal sections were stained for GFAP and were represented in the corresponding Paxinos atlas (Figure 2A). Confocal images from the sections were obtained using Zeiss 880 Airyscan microscope.

Behavioral modelling

In order to examine choices of mice in our task and formally define the computations that could underlie choices, we adopted a temporal difference reinforcement learning model which we developed previously and is set up to solve a partially observable Markov decision problem (Lak et al., 2017). Our 2AFC contrast detection task with unequal payoffs is an experimental realization of an environment in which, rather than simply finding itself in one of the two “left” or “right” states, the agent attributes a probability to each state, depending on the stimulus contrast. Stimulus contrast ranges from –0.5 to 0.5, where -0.5 and 0.5 indicates a stimulus with 50% contrast on the left or right side of the monitor, respectively.

In keeping with the standard psychophysical treatments of sensory noise, the model assumes that the internal estimate of the stimulus, $\hat{s}$, is normally distributed with constant variance around the true stimulus contrast: $p(\hat{s}|s) = \mathcal{N}(\hat{s}; s, \sigma^2)$. In the Bayesian view, the observer’s belief about the stimulus s is not limited to a single estimated value $\hat{s}$. Instead, $\hat{s}$ parameterizes a belief distribution over all possible values of s that are consistent with the sensory evidence. The optimal form for this belief distribution is given by Bayes rule:

$$p(s|\hat{s}) = \frac{p(\hat{s}|s).p(s)}{p(\hat{s})}$$

We assume that the prior belief about s is uniform, which implies that this optimal belief will also be Gaussian, with the same variance as the sensory noise distribution, and mean given by $\hat{s}$: $p(s|\hat{s}) = \mathcal{N}(s; \hat{s}, \sigma^2)$. From this, the agent computes a belief, i.e. the probability that the stimulus was indeed on the right side of the monitor, $p_R = p(s > 0 \mid \hat{s})$, according to:

$$p_R = \int_0^{\infty} p(s|\hat{s})$$

p_R represents the trial-by-trial probability of the stimulus being on the right side (and $p_L = 1 - p_R$ represents the probability of it being on the left).

The value of the two choices L and R can be computed as follows:

$$\begin{bmatrix} Q_R \\ Q_L \end{bmatrix} = \begin{bmatrix} q_{LR} & q_{RR} \\ q_{LL} & q_{RL} \end{bmatrix} \begin{bmatrix} p_L \\ p_R \end{bmatrix}$$

where q_{sa} represents the stored value of taking action *a* for stimulus *s*. For action selection, we used a ϵ-greedy rule which selects the action with higher value with probability of 1-T. T denotes small tendency in making random choices, which we represented by T_L and T_R (tendency to take left or right action, Figure 1), similar to lapse rate in standard psychometric methods. Using other choice functions such as softmax did not substantially changed our results. The outcome of this is thus the choice (L or R) and the expected value of the chosen option, Q_{Choice}, defined as:

$$Q_{Choice} = \begin{cases} Q_L \;\; if\ choice = L \\ Q_R \; if\ choice = R \end{cases}$$

Upon observing the stimulus and selecting a choice, the prediction error at the stimulus time is computable as:

$$\delta_{stimulus} = \; Q_{Choice} - V_{onset\ tone}$$

where $V_{onset\ tone}$ is the expected value of reward at the beginning of the trial, i.e. when the animal receives the auditory tone that the trial has started.

$$V_{onset\ tone} = \frac{q_{LL}+q_{LR} + q_{RL} + q_{RR}}{4}$$

After making the choice and receiving the reward, r, the reward prediction error is computed as:

$$\delta_{reward} = r - \; Q_{Choice}$$

Given this prediction error the value of stimulus-action pairs related to the taken choice, will be updated according to:

$$q_{La} \leftarrow q_{La} + \alpha . p_L . \delta_{reward}$$

$$q_{Ra} \leftarrow q_{Ra} + \alpha . p_R . \delta_{reward}$$

where α is the learning rate, and as before a is L or R.

Our experiments contained blocks of trials with unequal water reward or activation/suppression of VTA dopamine neurons. The payoff for incorrect choices was set to zero, and the value of smaller water reward was set to 1. Thus the payoff matrix for blocks with larger reward on the left or right, respectively, are:

$$outcome : \begin{bmatrix} 0 & 1 \\ 1 + x & 0 \end{bmatrix}, outcome : \begin{bmatrix} 0 & 1 + x \\ 1 & 0 \end{bmatrix}$$

where x represents the value of extra drop of water or optogenetic dopamine manipulation.

We fitted the model as well as reduced model variants on the choices of mice and cross-validated the necessity of model parameters (Figure S1D,E, Figure S4D,E). As described above, the full model included the following parameters: σ^2, T_L, T_R, x, α. Each reduced model did not include one of these parameters. For σ^2, one reduced model was set to have $\sigma^2 = 0$, representing a model with no sensory noise, and the other reduced model was set to have $\sigma^2 = \infty$, representing a model with extremely large sensory noise. In the reduced model examining the effect of T_L, T_R, we set both parameters to zero. For cross-validated fitting, we divided sessions of each mouse to 3 and performed a 3-fold cross validation. We performed the fit and parameter estimation on the training sessions and used the estimated parameters to predict choices in the test sessions and computed goodness of fit. For fitting, we performed exhaustive search in the parameter space expanding large value range for each of the parameters to find the best set of model parameters that account for the observed choices. To do so, we repeatedly fed the sequences of stimuli that each mouse experienced to the model, observed choices (iteration = 500), and computed the proportion that model made a leftward and rightward choice ($\hat{P}(L), \hat{P}(R)$) for each trial. We then calculated negative log likelihood as the average of $-\log\left(\hat{P}(\text{choice})\right)$, where $choice$ indicates the mouse's choice direction in each trial (Figure S1D).

Neuronal regression analysis

In order to quantify how each task event (stimulus, action, outcome) contributes to neuronal activity, and, the extent to which trial-by-trial variation in neuronal responses reflect animal's estimate of future reward and prediction error, we set up a neuronal response model (Figure S2C, S3A).

We modelled the spiking activity of a neuron during trial j, which we denote $R_j(t)$ as

$$R_j(t) = S_j K_s(t) * X^s{}_j(t) + A_j K_a(t) * X^a{}_j(t) + O_j K_o(t) * X^o{}_j(t)$$

In the above equation $K_s(t)$, $K_a(t)$ and $K_o(t)$ are the kernels representing the response to the visual stimulus, the action, and the outcome. $X^s{}_j(t)$, $X^a{}_j(t)$ and $X^o{}_j(t)$ are indicator functions which signify the time point at which the stimulus, action and outcome occurred during trial j. S_j, A_j and O_j are multiplicative coefficients which scale the corresponding kernel on each trial and * represents convolution. Therefore, the model represents neuronal responses as the sum of the convolution of each task event with a kernel corresponding to that event, which its size was scaled in each trial with a coefficient to optimally fit the observed response. Given the temporal variability of task events in different trials, the kernel for a particular task event reflects isolated average neuronal response to that event with minimal influence from nearby events. The coefficients provide trial-by-trial estimates of neuronal activity for each neuron.

The model was fit and cross-validated using an iterative procedure, where each iteration consisted of two steps. In the first step the coefficients S_j, A_j and O_j were kept fixed and the kernel shapes were fitted using linear regression. Kernels were fitted on 80% of trials and were then tested against the remaining 20% test trials (5-fold cross-validation). In the second step, the kernels were fixed and the coefficients that optimized the fit to experimental data were calculated, using linear regression as well. Five iterations were performed. In the first iteration, the coefficients were initialized with value of 1. We applied the same analysis on the GCaMP responses (Figure 3, S3).

We defined the duration of each kernel to capture the neuronal responses prior or after that event and selected longer kernel duration for the GCaMP data to account for Ca^{+2} transients (PL spike data: stimulus kernel: 0 to 0.6 s, action kernel: -0.4 to 0.2 s, outcome kernel: 0 to 0.6 s; GCaMP data: stimulus kernel: 0 to 2 s, action kernel: -1 to 0.2 s, outcome kernel: 0 to 3s, where in all cases 0 was the onset of the event). For both spiking and GCaMP data, the neuronal responses were averaged using temporal window of 20 and 50 ms, respectively, and were then z-scored.