Dopaminergic and prefrontal dynamics co-determine mouse decisions in a spatial gambling task

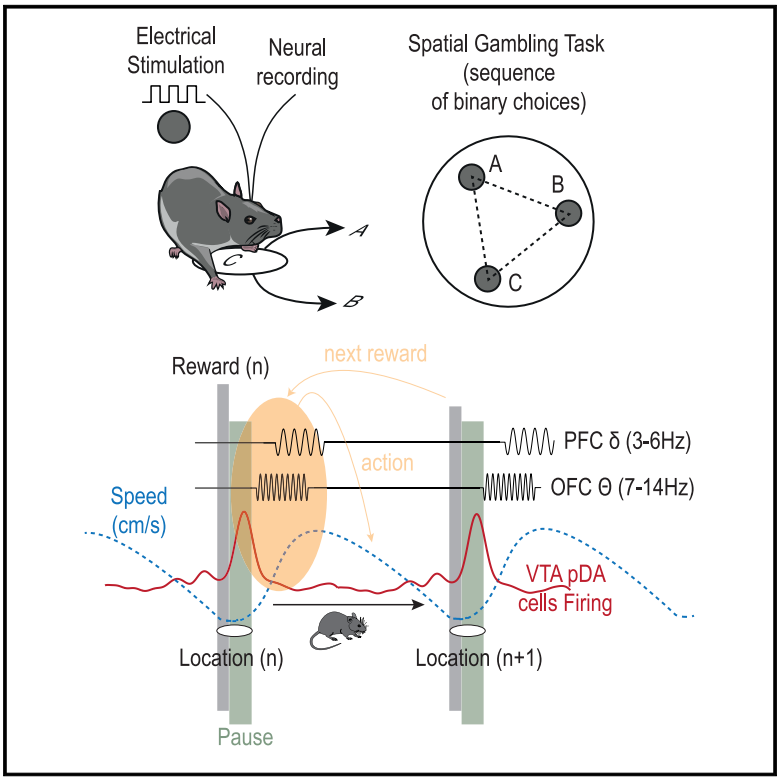
Elise Bousseyrol, Steve Didienne, Samir Takillah, Clement Prevost-Solié, Maxime Come, Tarek Ahmed Yahia, Sarah Mondoloni, Eléonore Vicq, Ludovic Tricoire, Alexandre Mourot, et al.

Cell Reports

Article

Dopaminergic and prefrontal dynamics co-determine mouse decisions in a spatial gambling task

Graphical abstract



Highlights

- Self-paced actions arise from contextual reorganization of mesocortical dynamics
- VTA, PFC, and OFC complementarily encode predictions and errors about outcomes
- Distributed "firing then oscillations" dynamics set the goal, initiation, and pace of actions
- VTA and PFC antagonistically promote and inhibit motivation by reward uncertainty

Authors

Elise Bousseyrol, Steve Didienne, Samir Takillah, ..., Alexandre Mourot, Jérémie Naudé, Philippe Faure

Correspondence

jeremie.naude@igf.cnrs.fr (J.N.), phfaure@gmail.com (P.F.)

In brief

The neural mechanisms by which animals initiate goal-directed behaviors, choose between alternatives, and explore potentially informative ones are far from understood. Using a mouse gambling task, Bousseyrol et al. identified a sequence of oscillations and firings in ventral tegmental, orbitofrontal, and prefrontal areas that jointly encode choice and action.

Bousseyrol et al., 2023, Cell Reports *42*, 112523
May 30, 2023
https://doi.org/10.1016/j.celrep.2023.112523

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Dopaminergic and prefrontal dynamics co-determine mouse decisions in a spatial gambling task

Elise Bousseyrol,[1,2,4] Steve Didienne,[1,2,4] Samir Takillah,[1,2] Clement Prevost-Solié,[1,2] Maxime Come,[1,2] Tarek Ahmed Yahia,[1] Sarah Mondoloni,[1] Eléonore Vicq,[1] Ludovic Tricoire,[1] Alexandre Mourot,[1,2] Jérémie Naudé,[1,3,5,*] and Philippe Faure[1,2,5,6,*]

[1]Sorbonne Université, INSERM, CNRS, Neuroscience Paris Seine – Institut de Biologie Paris Seine (NPS – IBPS), 75005 Paris, France
[2]Brain Plasticity Laboratory, CNRS, ESPCI Paris, PSL Research University, 75005 Paris, France
[3]CNRS, Université de Montpellier, INSERM – Institut de Génomique Fonctionnelle, 34094 Montpellier, France
[4]These authors contributed equally
[5]These authors contributed equally
[6]Lead contact
*Correspondence: jeremie.naude@igf.cnrs.fr (J.N.), phfaure@gmail.com (P.F.)
https://doi.org/10.1016/j.celrep.2023.112523

SUMMARY

The neural mechanisms by which animals initiate goal-directed actions, choose between options, or explore opportunities remain unknown. Here, we develop a spatial gambling task in which mice, to obtain intracranial self-stimulation rewards, self-determine the initiation, direction, vigor, and pace of their actions based on their knowledge of the outcomes. Using electrophysiological recordings, pharmacology, and optogenetics, we identify a sequence of oscillations and firings in the ventral tegmental area (VTA), orbitofrontal cortex (OFC), and prefrontal cortex (PFC) that co-encodes and co-determines self-initiation and choices. This sequence appeared with learning as an uncued realignment of spontaneous dynamics. Interactions between the structures varied with the reward context, particularly the uncertainty associated with the different options. We suggest that self-generated choices arise from a distributed circuit based on an OFC-VTA core determining whether to wait for or initiate actions, while the PFC is specifically engaged by reward uncertainty in action selection and pace.

INTRODUCTION

Animals often base decisions on internal representations of their goals.[1,2] This includes when to act but also which option to take. When faced with several alternatives, they do not always exploit the option with the highest reward expectation but instead explore less rewarded options to gain information.[3,4] The neural mechanisms by which animals initiate goal-directed action, decide between actions, and explore potentially informative action are far from being understood. The computational theory of reinforcement learning identifies phasic dopamine (DA) release with a reward prediction error (RPE); i.e., the comparison between actual and expected reward. The DA RPE would constitute a teaching signal for learning appropriate stimulus-action responses.[5,6] In this framework, cortices would provide subcortical areas with information about the current state and environmental options, and the basal ganglia would select among these options to initiate the goal-directed action.[2,7] Hence, in these theories, frontal cortices and DA only indirectly affect decisions by affecting subsequent trials of a task rather than the current one.[8,9]

However, the basal ganglia might not be the only locus of action selection.[10] DA and frontal areas have also been described as having more direct, online roles in decision-making. Good-based models place the choice or comparison processes at the level of the orbitofrontal cortex (OFC) and prefrontal cortex (PFC).[11–14] Theories of cognitive control assign a top-down, potentially inhibitory role to the OFC and PFC through computing goals, plans, and task rules; i.e., higher-order decision-making.[15] These different accounts all point at a direct, active role of frontal areas in ongoing choices. Phasic DA also influences ongoing behavior and self-paced actions by modulating the vigor of actions leading to rewards.[16,17] Mixed results have been obtained on DA facilitating action initiation itself, depending on the type of task, DA nuclei, and intensity of DA manipulation.[18–21] Mesolimbic DA (particularly in the nucleus accumbens) has a major role in locomotion and reward,[17,22] but the effects of DA interaction with the broader decision circuitry on ongoing motivation remain unclear. Given the respective roles of DA and frontal cortices in value-based decision making and action initiation, the question of the coordination between nodes of this mesocortical circuit in learning processes, decision-making and exploration, and motor execution arises.

Here we used an experimental paradigm where mice perform a sequence of choices to obtain rewards associated with intracranial self-stimulation.[23–25] Mouse behavior displayed hallmarks of self-paced action, with initiation, pace, and decisions underdetermined by environmental information but influenced by internal representations of the reward

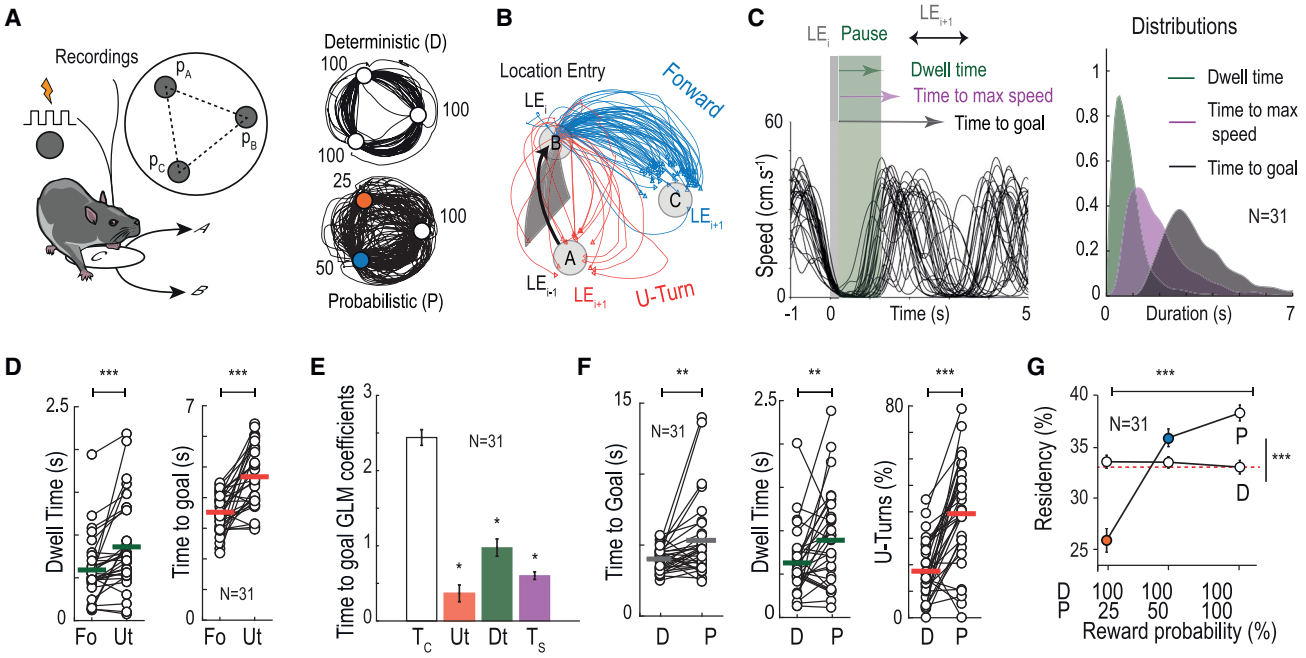


Figure 1. Self-paced decisions in a mouse spatial task based on ICSS

(A) Left: task design. Three explicit locations in an open field are associated with a given probability of intracranial self-stimulation (ICSS) delivery (p = 100%, 50%, 25%) when the mouse is detected in the location area. Animals could not receive two consecutive ICSSs at the same location. Right: examples of trajectories (5 min) showing that mice alternated between rewarding locations in the deterministic (D) and probabilistic (P) contexts.

(B) Animals varied between forward (Fo) trajectories, in which mice keep the direction of their last choice, and U-turn (Ut) trajectories, in which mice went backd to their previous location.

(C) Left: example of instantaneous speed profiles after learning in the D context, showing that animals almost stopped at the ICSS time, upon location entry (location entry [LE]), stayed immobile for a short dwelling period, then accelerated toward their next location. These bouts of activity can be described using three observables: the dwell time (speed < 10 cm/s), time to maximal speed, and total time to goal from one location to the next. Right: distribution for these three parameters, for all trials of all mice, at the end of the D context.

(D) In the D context, dwell time and time to goal were higher when mice performed Ut compared with Fo trajectories (dwell time: $\Delta = 0.27$ s, paired two-sided Wilcoxon signed-rank test, $W_{(30)} = -38$, $p < 0.001$; time to goal: $\Delta = 1.16$ s, paired Student's t test, $T_{(30)} = -4.95$, $p < 0.001$). Colored horizontal bars represent the means.

(E) Coefficients of the generalized linear model (GLM) of time to goal in the D context: constant term (Tc), Ut (Ut or Fo) (Student's t test, $T_{(30)} = 3.31$, $p = 0.0024$), dwell time (Dt) (two-sided Wilcoxon signed-rank test, $W_{(30)} = 496$, $p < 0.001$), and time to maximal speed (Ts) (Student's t test, $T_{(30)} = 11.76$, $p < 0.001$). Vertical bars represent SEM. Asterisks indicate a significant impact on the time to goal.

(F) Time to goal, Dt, and proportion of Uts increased in the P context compared with the D context (time to goal: paired two-sided Wilcoxon signed-rank test, $\Delta = 1.14$ s, $W_{(30)} = 104$, $p = 0.005$; Dt for all trajectories: paired Student's t test, $T_{(30)} = -2.64$, $p = 0.01$, $\Delta = 0.26$ s; Dt for Fo trajectories only: Student's t test, $T_{(30)} = -3.27$, $p = 0.0027$, for time to goal; Ut: paired Student's t test, $\Delta = 0.20$ s, $T_{(30)} = -4.64$, $p < 0.001$). Colored horizontal bars represent the means.

(G) Proportion of choices of the three rewarding locations as a function of reward probability in the P and D contexts. Effect of the P context on choice distribution: $F_{(30,2)} = 48.5$, $p < 0.001$; same for D context: $F_{(30,2)} = 0.2$, $p = 0.81$, one-way ANOVA. Effect of probabilities on choices in the P context: $F_{(1,2)} = 31.8$, $p < 0.001$, two-way ANOVA. Vertical bars represent SEM.

***p < 0.001,**p < 0.01,*p < 0.05.

context.[1,22] We show that the ventral tegmental area (VTA), PFC, and OFC coordinated their activities into a sequence of distributed firing and oscillations. Such a sequence emerged with learning as a reorganization of existing dynamics. Combining electrophysiology with causal manipulations, we unveil that OFC-VTA interactions set self-initiation in every reward context, with more specific PFC involvement in decisions under uncertainty and exploration. The VTA and PFC can act synergistically to self-pace the actions but may have antagonistic roles in pondering the influence of uncertainty on choices, particularly for exploration. Our study highlights how the mesocortical circuit self-generates decisions through distributed but distinct computations.

RESULTS

Mouse actions underdetermined by stimulus cues, but shaped by internal representations, indicate self-paced decisions

We used a spatial version of a bandit task adapted to mice.[23–25] Three equidistant locations explicitly marked in an open field were associated with rewards delivered as intracranial self-stimulation (ICSS) in the medial forebrain bundle (MFB)[26] (Figure 1A, left). Mice could not receive two consecutive ICSS at the same location; they therefore alternated between rewarding locations (Figure 1A, right). The task is thus a sequence of movements (i.e., trials) between rewarded locations. There is no stimulus cue to

trigger the movement or to specify the direction or initiation of each trial, but the animal can use environmental and explicit cues on the floor to locate the position of the targets and guide its behavior. Mice were initially trained in a deterministic (D) context in which all locations were associated with a certain ICSS delivery (p = 100%). Then, mice were subjected to a probabilistic (P) context in which each location was associated with a different probability of ICSS delivery (p = 100%, 50%, and 25%; Figure 1A) to assess how self-paced actions and meso-cortical representations depend on outcome expectations. Their behavior in the task after learning in the D context consisted of sequences of trials in between rewarding locations. First, on each rewarded location, mice could either circle forward along the three locations or perform a U-turn to come back to the previous (i-1) location (Figure 1B). Second, a trial was characterized by a "template" bout of locomotion: a movement initiation toward the next location, acceleration followed by deceleration, and a pause at the next location (Figure 1C, left). Although the ICSS itself caused a decrease in the velocity (random ICSS in the home cage; Figure S1A), this effect was small compared with the total restructuring of the locomotion pattern in the task. Execution of this ballistic velocity profile was characterized by an important trial-to-trial variability, notably regarding the time during which animals dwelled at a rewarded location before initiating a new trial, the time to reach maximal speed, and the overall time to goal (from one location entry to the next; Figure 1C, right). Because these successive timings were nested within the behavioral sequence, their variability could correspond to a global decision on the overall timing of the trial. In such case, variability of intermediate timings would be correlated either positively (with the first timings already explaining the subsequent variability) or negatively (with the variability in each timing compensating in a zero-sum fashion). Alternatively, successive, independent addition of variability at each stage of the trial could signal distinct decisions about movement direction, initiation, and vigor. In the D context, the direction (U-turn or forward) impacted the dwell duration (an early trial timing) and the time to goal (Figure 1D). Multiple linear regressions (with orthogonalized predictors, model $p < 0.05$ for every animal; STAR Methods) showed that the trajectory direction, dwell time, and time to maximal speed all had (independent from each other) a significant impact on the time to goal (Figure 1E). Hence, each stage of a trial presented a successive addition of independent variability, suggesting that trajectory direction, initiation, and vigor all constituted decisions for the animals.[1,2]

Self-paced decisions depend on the expected value of potential outcomes.[1,22] We thus assessed whether the timings of the trial were affected by the animals' internal estimates of the potential rewards by comparing behaviors in the D and P context (with different probabilities of ICSS delivery at each location; p = 100%, 50%. and 25%; Figure 1A). After training in the P context, the time to goal and dwell time increased compared with the D context, even when considering only forward trajectories, suggesting decreased motivation because of a decrease in reward frequency in the P context (Figure 1F). The proportion of U-turns increased as well, reflecting an adjustment of the trajectory directions to the respective payoffs of the locations; while mice visited the three locations uniformly in the D context, in the P context, they more often visited the locations associated with the highest ICSS probabilities (Figure 1G). Hence, mouse behavior displayed hallmarks of self-paced decisions; the direction, initiation, and vigor of actions presented independent variability and were not determined by a stimulus cue but, rather, were influenced by the potential outcomes of the actions.

The distributed mesocortical sequence associated with self-generated decisions is reward context dependent

We next characterized the mesocortical dynamics associated with self-paced decisions. We recorded from putative DA neurons (pDAns; n = 136 neurons) in the VTA from wild-type (WT) mice (n = 12 mice) using extracellular multielectrodes (Figure 2A, left). All neurons met the electrophysiological and pharmacological criteria used to identify DA cells *in vivo*[27,28] (Figure S1; STAR Methods). In another group of mice, bipolar electrodes were also chronically implanted bilaterally into the PFC and OFC (n = 23 mice), allowing us to record extracellular field potentials (EFP) and population spiking activity (Figure 2A, right; Figures S1 and S2; STAR Methods).

At the end of the D context, pDAn displayed a transient increase in firing frequency early in the trial (hereafter called "early phasic" activity), after location entry and ICSS delivery, during the dwelling period (Figure 2B; n = 65 cells). The OFC was characterized by an increase in firing before the reward, followed by a dip during the ICSS (potentially because of the stimulation masking spikes; hence, it was not analyzed further) and then a rebound that was concurrent with pDAn early phasic activity (Figure 2C, top). After the dwell time, when animals started to move toward the next location (0.5- to 1.5-s window), oscillation power in the θ band (7–14 Hz) increased in the OFC EFP (Figure 2C, bottom). In contrast, no specific activity was observed in the PFC around the location entry, neither in population spiking nor in EFP (Figure 2D; as with the OFC, the activity during the ICSS was not analyzed because of potential masking). Consistent with this temporal order, firing and θ oscillations in the OFC displayed a lagged cross-correlation (Figure 2E), indicating a transition from increased spiking to θ oscillations around the time of self-initiation of the trial by the mice. Although some of the oscillations in the PFC and OFC co-occurred, we did not observe any increase in coherence in either context (Figure S3).

We next asked how mesocortical activity changed with the reorganization of animals' choices when faced with the uncertainty of P reward delivery. pDAn early phasic activity was still present at the end of the P context, and its amplitude was not different from that in the D context (Figure 2F). However, different trials in the P context correspond to distinct situations (reward omissions, reward expectations, etc.) that are analyzed further down (Figures 4 and 5). The increase in OFC firing during dwelling and latter power increase in θ oscillations were similar in both contexts (Figure 2G). By contrast, in the P context, an increase in population firing (0.5- to 1-s window after location entry, post omission only) and δ (3–6 Hz) oscillation power (0.5- to 1.5-s window after location entry) emerged in the PFC but only when ICSS was omitted (Figure 2H). No preferential temporal lag was observed between PFC firing and δ oscillations after location entry (Figure 2I), suggesting no clear temporal order between firing and oscillations. Overall, the mesocortical network, classically

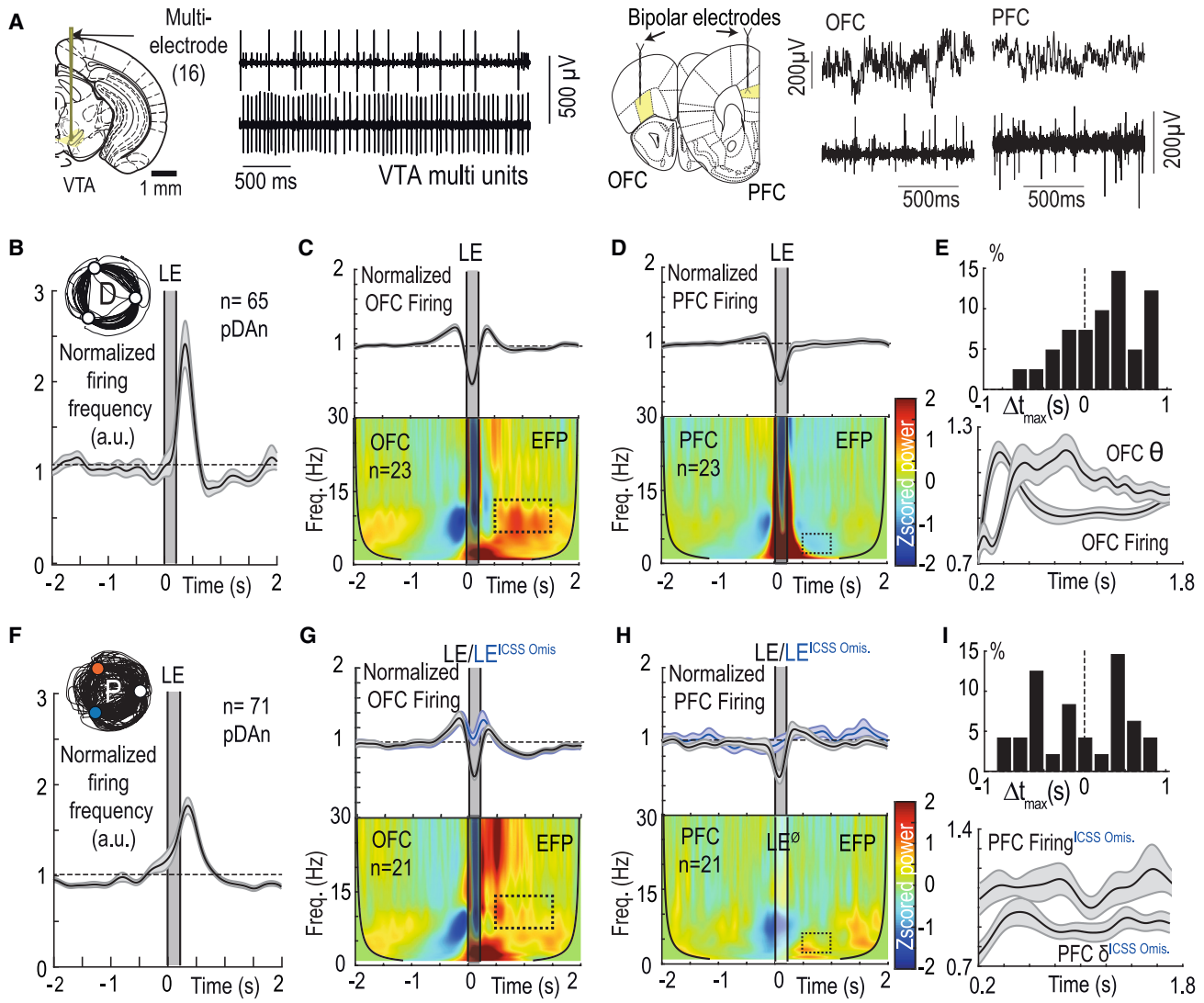


Figure 2. VTA, PFC, and OFC during D and P reward contexts

(A) Left: filtered (600–6000 Hz) extracellular recordings in the VTA show multiple unit activities. Right: extracellular recordings in the OFC and PFC show extracellular field potential (EFP; filtered 0.1–300 Hz) (top) or population activity (filtered 600–6000 Hz) (bottom).

(B) Normalized firing frequency from VTA pDAns around the LE; i.e., reward delivery in the D context (pDAn mean firing over 0.3–0.8 s is increased from baseline; one-sample Wilcoxon test, $W_{(64)} = 1{,}572$, $p = 0.001$). Data are presented as mean ± SEM across units.

(C) Top: normalized firing frequency from the OFC population (mean ± SEM across units) around LE in the D context (i.e., reward delivery) (OFC mean firing over 0.3–0.8 s, Student's t test, $T_{(39)} = 2.57$, $p = 0.01$). Bottom: 0- to 30-Hz range time-resolved power spectral density (PSD; using a complex Morlet wavelet transform) of OFC EFP around reward delivery. PSD is *Z* scored over the 2-s period preceding the LE (mean OFC θ 7- to 14-Hz power over 0.5–1.5 s, Student's t test, $T_{(22)} = 5.12$, $p < 0.001$ across units).

(D) Same as (C) for the PFC.

(E) Top: time lag between the maximal OFC θ oscillation power and the maximal OFC firing (time of the maximum θ power minus time of maximum firing: two-sided Wilcoxon-Mann-Whitney test, $U_{(78)} = 1{,}371$, $p = 0.002$). Bottom: superposition of OFC θ oscillation and population firing frequency. Data are presented as mean ± SEM across units.

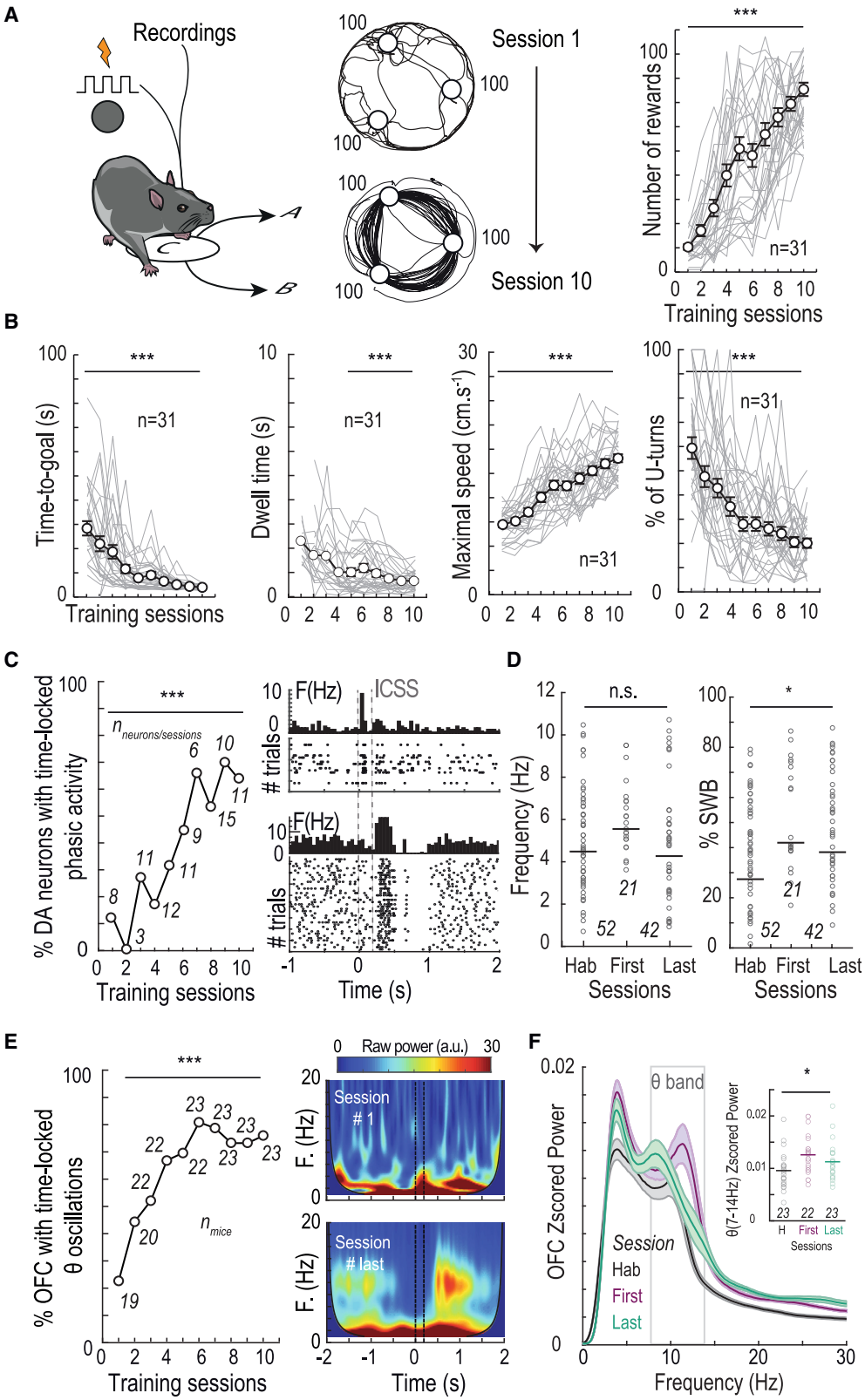
(F–H) Same as (B)–(D) but in the P context. PFC and OFC firing are also given for LE (black) and $LE^{ICSS\ Omission}$ (blue) conditions.

(F) Mean pDAn firing frequency is increased (from baseline) over 0.3–0.8 s: one-sample t test, $T_{(74)} = 7.61$, $p < 0.001$. Difference compared with the D setting shown in (B): two-sided Wilcoxon-Mann-Whitney test, $U_{(64)} = 4{,}345$, $p = 0.32$.

(G) Difference between P and D settings (all trials): Top: OFC mean firing frequency during 0.3–0.8 s: Student's t test, $T_{(91)} = 0.56$, $p = 0.6$. Bottom: OFC mean θ power during 0.5–1.5 s: Student's t test, $T_{(42)} = 1.12$, $p = 0.3$.

(H) Top:) PFC mean firing over 0.5–1 s post omission: two-sided Wilcoxon signed-rank test, $W_{(49)} = 866$, $p = 0.03$. Bottom: the PFC PSD is centered on reward omissions ($LE^{ICSS\ Omis.}$) to prevent the ICSS artifacts from obscuring the low frequency power. PFC δ mean power during a 0.5- 1-s window post omission: Student's t test, $T_{(20)} = 2.07$, $p = 0.05$; difference with the D context: two-sided Wilcoxon-Mann-Whitney test, $U_{(20)} = 355$, $p < 0.001$.

(I) Top: time lag between the maximal PFC δ oscillation power and the maximal PFC firing, post omission only (time of the maximum δ power minus time of maximum firing: paired two-sided Wilcoxon signed-rank test, $W_{(47)} = 2{,}121$). Bottom: superposition of PFC δ oscillation power and population firing frequency.



(legend on next page)

associated with stimulus-triggered decision-making,[8,29,30] was also recruited during self-paced decisions with or without engagement of the PFC, depending on the context (P vs. D).[31]

A self-generated mesocortical sequence emerges with learning as a reorganization of existing dynamics

The influence of reward context on VTA, PFC, and OFC activities suggests that the mesocortical dynamics associated with self-paced actions reorganize in response to the outcomes of these actions. We thus investigated the emergence of these dynamics with learning in naive mice in the D context (Figure 3A, left and center). The number of reward locations visited per session increased throughout the sessions (Figure 3A, right), confirming place reinforcement. The time to goal (Figure 3B, left) and dwell time (Figure 3B, center left) decreased accordingly, while maximal speed increased (Figure 3B, center right). Finally, the proportion of U-turns decreased (Figure 3B, right), indicating that mice learned to optimize their trajectories and reduced the motor cost associated with U-turns.[23]

This behavioral learning was associated with a modification of VTA pDAn dynamics. The proportion of pDAns with a firing rate significantly higher than baseline (STAR Methods) increased with the training sessions (Figure 3C, left). Early in learning, when the behavior is still dominated by spontaneous locomotion, phasic pDAn firing was sometimes observed at the time of the ICSS (Figure 3C, top right), but not in every neuron (Figure S4A), and was never observed during the dwell time. In contrast, at the end of learning, increased pDAn activity appeared at a specific time point during the dwell time (Figure 3C, bottom right). The time-locked increase in pDAn firing and the increase in the number of trials with learning did not result in an overall (session-wide) increase in pDAn firing frequency (Figure 3D, left) nor in a shift of pDAn firing pattern toward increased bursting overall[32,33] (i.e., there was an increase from naive mice to first sessions but not from first session to last session), as estimated with the percentage of spikes within bursts (%SWB; STAR Methods; Figure 3D, right). Hence, early phasic activity in DA neurons did not rely on additional spikes or on increased burstiness but, rather, on dynamic re-organization of spikes toward bursting activity at behaviorally relevant times (Figures S4B and S4C). Similarly, the total number of OFC EFPs in which time-locked θ oscillation power was higher than baseline increased with learning (Figure 3E, left; Figure S4D). These θ oscillations emerged consistently around the same time; i.e. just after the end of the dwell time, after learning (Figure 3E, right; Figure S4E). Compared with power spectra from naive animals undergoing a habituation (Hab) session (Figure 3F) to the open field, total θ power increased in the first and last sessions of the D context compared with naive animals, but no difference was observed, on average, between the first and last sessions. Hence, θ oscillations increased early in learning and reorganized to time lock at mouse departure while learning progressed. By contrast, no change in PFC oscillatory activity was observed throughout learning, as found at the end of the D context (Figure 2). Overall, during learning in the D context, optimization of trajectories and speed profiles was associated with reorganization of VTA DAns firing toward time-locked bursting during dwell time and of OFC activity toward time-locked θ oscillations when animals accelerate.

Early phasic activity of VTA DA cells, together with frontal firing and oscillations, forms a distributed signal for outcome discrepancy and expectations

Learning theories propose that behavior, and underlying brain activity, reorganize when unexpected outcomes occur; the discrepancy between received and expected reward (the RPE) may be used by animals to update their internal representations.[34–36] This RPE is thought to be computed in mesocorticolimbic areas,[9,37,38] most notably by VTA DA cells,[6] but signals related to evaluation of outcomes are also present in the frontal cortex.[39] Alternative interpretations, such as signaling outcome expectancy or prediction, may account for observed increases in neural activity associated with reward delivery.[40] We thus evaluated VTA, OFC, and PFC activities during expected rewards, unexpected rewards, and omissions. This allowed us to disentangle the direct effect of the ICSS on mesocortical activity from the learned reward representation. Early bursting phasic

Figure 3. Early VTA and OFC activities emerge with learning

(A) Left: schematic of learning in the D context with trajectory examples from session 1 (top center) and session 10 (bottom center). Right: number of rewards along learning (repeated-measures ANOVA, $F_{(9,30)} = 62.6$, $p < 0.001$). Dots and vertical bar are mean ± SEM across subjects. Gray lines indicate modifications of the number of rewards per individual (n = 31 animals).

(B) Same as (A), from left to right: time to goal (repeated-measures ANOVA, $F_{(9,30)} = 27$, $p < 0.001$), Dt (only defined for sessions 5–10, repeated-measures ANOVA, $F_{(5,30)} = 7$, $p < 0.001$), maximal speed (repeated-measures ANOVA, $F_{(9,30)} = 48.6$, $p < 0.001$), and proportion of Uts (repeated-measures ANOVA, $F_{(9,30)} = 20.7$, $p < 0.001$) along the learning sessions. Dots and vertical bar are mean ± SEM.

(C) Left: proportion of pDAns modulated in between two locations along learning sessions (n = 10 sessions) (χ^2 test, $\chi^2 = 300$, $p < 0.001$). Right: examples of raster plots, centered on LE, for a VTA pDAn early in learning (top) and another at the end of learning (bottom).

(D) Average firing frequency (left) and % spikes within bursts (%SWB, right) throughout learning in the D context (Hab: open field without ICSS prior to the learning stage; first: sessions 1–5; last: sessions 6–10). Modification of firing frequency, ANOVA, $F_{(2,115)} = 1.8$, $p = 0.18$; same for %SWB: ANOVA, $F_{(2,115)} = 3.12$, $p = 0.048$. Horizontal bars represent the means.

(E) Left: proportion of OFC θ power (7–14 Hz) modulated in between two locations along learning sessions (χ^2 test, $\chi^2 = 44.5$, $p < 0.001$ across units). Right: examples of OFC time-resolved PSD centered on LE for one OFC early in learning (top) and for the same OFC at the end of learning (bottom).

(F) Mean *Z*-scored power of Fourier transform spectra of OFC during open-field habituation (Hab; black, n = 21; STAR Methods), early in learning (Det first, purple, n = 20) and at the end of the learning (Det end, green, n = 19). Data are presented as mean ± SEM across subjects. Power in the θ-band frequency (7–14 Hz, gray box) shows a significant difference between the three conditions (ANOVA, $F_{(2,65)} = 3.94$, $p = 0.024$), with differences between the Hab session and the first (Student's t test, $T_{(43)} = 2.80$, $p = 0.015$, $\Delta = +0.003$) and last (paired Student's t test, $T_{(22)} = -2.98$, $p = 0.02$, $\Delta = +0.002$) sessions of the D context. There were no difference between first and last sessions (Student's t test, $T_{(43)} = 1.22$, $p = 0.23$).

***$p < 0.001$,**$p < 0.01$,*$p < 0.05$, n.s. $p > 0.05$.

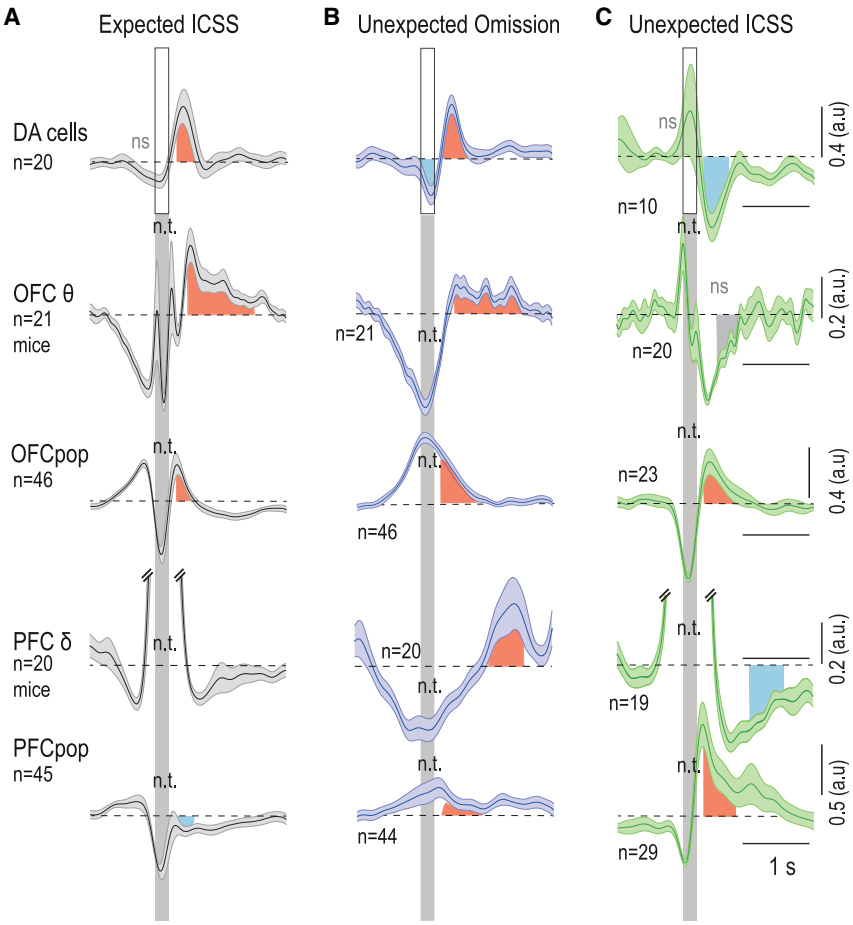


Figure 4. VTA, OFC, and PFC activities following expected reward, unexpected omission, and unexpected reward

(A) From top to bottom: VTA pDAn normalized firing, OFC θ oscillation power, OFC normalized population firing, PFC δ oscillation power, and PFC normalized population firing, centered on expected reward delivery upon LE at the beginning of the P context. Data are presented as mean ± SEM across units (for spiking) and across subjects (for oscillations). ICSS artifacts masked spikes from OFC and PFC population firing and induced a deflection in OFC and PFC oscillations, so the time periods corresponding to ICSS duration were not analyzed for these signals (gray shade; n.t. indicates that the time point was not tested).

(B) Same as (A), centered on unexpected omission of reward delivery upon LE at the beginning of the P context.

(C) Same as (A), centered on unexpected reward delivery upon random ICSS in the home cage, before the beginning of the conditioning. Red, significant increases; gray, ns activity; light blue, significant decreases. The timescales and ordinates are indicated for each line on the right and are identical for the three conditions shown on the same line.

n.s. $p>0.05$

activity in pDAns at the beginning of the P context occurred during the dwelling period (Figure 4A, top row; Figure S5A), as described previously for the end of the D and P contexts (Figure 2). Importantly, this increased DA activity was observed even when the ICSS reward was unexpectedly omitted at the beginning of the P context (Figure 4B, top row; Figure S5A) but did not appear, on average, after an unexpected, random stimulation reward in the home cage (Figure 4C, top row; Figure S5A, but see Figure S4F for examples of pDAns responding or not responding to random stimulation). This is a clear indication that early phasic activity at the beginning of the trial is not generated as a response to the previous stimulation reward. Rather, it is related to expectation of the upcoming reward.[35,41] In this framework, DA cells not only signal the difference between actual and expected reward but also integrate the (discounted) expectation of future rewards predicted by the current state and actions of the animal.[6] Our results are thus consistent with pDANs computing a temporal difference (TD) RPE at two characteristic time points: during the ICSS (or at the omission time) and during the early phasic activity (prediction of the next ICSS). Alternatively, early phasic activity may also relate to invigoration of the next movement.[20,21] The decreased pDAn activity observed right after random stimulation in the home cage further suggests that early phasic activity in pDANs occurs specifically upon learning of the self-paced actions. Moreover, pDAn activity decreased at the time of the expected ICSS (but not at the time of the early phasic activity) during unexpected omissions (compared with baseline and with the expected ICSS condition; Figure 4B, top row; Figure S5A, right), consistent with pDANs computing a TD RPE at the time of the ICSS and at the time of the early phasic activity.[5,35,42]

Likewise, in the OFC, neither the increased power of θ oscillations when mice accelerate nor the increased firing during dwell time were triggered by the previous stimulation reward (Figure 4A; Figure S5B). Indeed, both were observed during omission trials, and no difference was observed between the ICSS and omission conditions (Figure 4B; Figure S5B). Furthermore, unexpected, random stimulation rewards did not generate θ oscillations (Figure 4C; Figure S5B), indicating specific involvement of the OFC in active behavior. An unexpected reward increased OFC population firing (Figure 4C; Figure S5B, right), suggesting an influence of expected and unexpected outcomes on OFC firing activity.

Finally, in the PFC (Figure 4, fourth and last rows), unexpected reward omission induced δ oscillations and increased population firing (Figure 4B; Figure S5C). Hence, δ oscillations and increased firing in the PFC, which were observed in the P context (Figure 2) but not in the D context (Figure 4A), were already present in the first omission trials. By contrast, unexpected stimulation reward decreased δ oscillations (Figure 4C, fourth row; Figure S5C) and increased population firing (Figure 4C, last row; Figure S5C), suggesting involvement of the PFC specifically following errors; i.e., unexpected outcomes (either reward or omission).

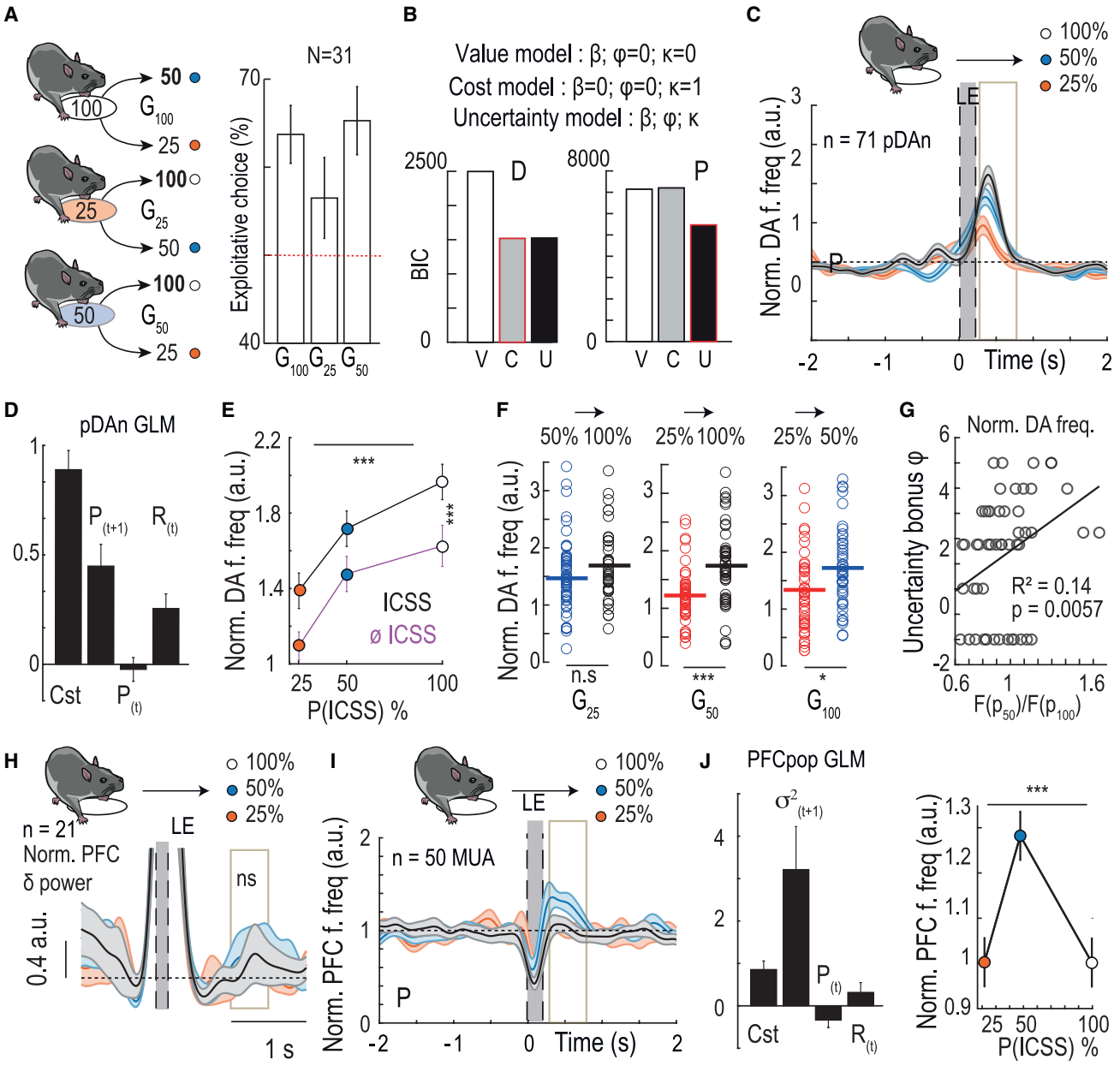


Figure 5. Complementary encoding of choices, value, and cost by the VTA, OFC, and PFC

(A) Left: schematic of the three Gs mice are facing in the P context. Right: proportion of choices for the location associated with the highest reward probability for each G (G_{25}: 100% vs. 50%, G_{50}: 100% vs. 25% and G_{100}: 50% vs. 25%) at the end of the P context (n = 31 mice). Vertical bars represent SEM across subjects.

(B) Bayesian information criteria (BICs) computed using three models of choice selection at the end of the D context (left) and of the P context (right). Red boxes surround the smaller BIC, indicating the best fit. V model: softmax with β only (value sensitivity model), C model: softmax with κ only (κ model), U model: softmax with β, κ, and φ (φ model) (STAR Methods).

(C) Normalized firing frequency (a.u.) of pDAns at the end of the P context, centered on LEs. Trials are sorted according to the next chosen location. A gray box indicates the quantification window.

(D) Coefficients of the GLM of VTA pDAn firing in the P context: Tc, reward probability of the goal (P_{t+1}), reward probability of current location (P_t), and outcome delivered at current location (categorical variable, ICSS or omission). Vertical bars represent SEM across units.

(E) Quantification of pDAn firing frequency according to the reward probability of the goal (ANOVA, $F_{(2,291)} = 16.81$, p = 0) when animals received the ICSS or not (Ø, purple, ANOVA, $F_{(1,291)} = 13.76$, p = 0.0002). Data are presented as mean ± SEM across units.

(F) Quantification of pDAn firing frequency according to the reward probability of the goal in the different Gs (G_{25}: 100% vs. 50%, G_{50}: 100% vs. 25%, and G_{100}: 50% vs. 25%). Horizontal bars represent the means.

(G) Phasic encoding of uncertainty by pDAns (activity related to 50%, p_{50}, versus 100%, p_{100}, reward probability of the chosen locations) against φ from the model (R^2 = 0.14, p = 0.006).

(legend continued on next page)

Overall, we observed, in VTA, OFC, and PFC firing and oscillations, distributed encoding of errors related to unexpected rewards and omissions, which is likely used for behavioral learning. Furthermore, we ruled out that the mesocortical dynamics observed during the dwell time were triggered by the preceding ICSS because they were observed after omission but not after random ICSS.

Distributed and complementary representations of decision parameters in mesocortical structures

Because the mesocortical dynamics associated with decisions were not solely caused by the immediately preceding reward, we searched for a neural signature of the upcoming outcome, which is thought to guide self-paced decisions.[1,2] By modeling how choices depended on the reward probability and U-turn motor requirements, we determined which task parameters actually affected animals' choices.

Because mice could not receive two consecutive rewards at the same location, they had to choose between the two remaining locations (Figure 5A, left). We expressed this succession of binary choices in the P context as the proportion of exploitative choices (i.e., option with the highest reward probability) in the following three gambles (Gs): G_{25} (100% vs. 50%), G_{50} (100% vs. 25%), and G_{100} (50% vs. 25%) reward probabilities (Figure 5A). Mice displayed a preference for the highest reward probability in G_{100} ($p = 0.02$) and G_{50} ($p < 0.001$) but not in G_{25}, in which they equally chose the locations associated with 50% and 100% reward ($p = 0.20$). This replicates our previous studies and can be explained by mice assigning a positive value to uncertainty, which is zero for predictable outcomes (here, 100% probability) and maximal for the most unpredictable outcome (50% probability).[24,25] We used a model-based analysis to disentangle the influence of expected reward and uncertainty (variance), which co-vary in this setup, on choices. Because choice behavior may differ between mice (Figures S6A and S6B), we modeled individual data using alternative models of decision-making.[24] Value-based models implementing the hypothesis that mice are guided by their outcome expectations explained the choices better than models corresponding to random choices (Figure S6C). Compared with the D context, in which choices could be explained by the motor cost (negative value of U-turns [κ], favoring forward trajectories), in the P context, animals added an uncertainty bonus (φ) to the expected reward as a total positive value, discounted by κ (Figure 5B). Model comparison[24] (STAR Methods; Figure S6C) confirmed that φ explained choices better than a saturating value function (in which value stops increasing with reward probabilities above 50%). Because animals choose according to outcome properties, this further suggests that behavior is goal directed in this task.[22,24]

We thus assessed the encoding of expected reward, uncertainty, and κ in mesocortical activity. We did not find any encoding of κ in any of the recordings (Figure S7A). By contrast, VTA pDAn activity scaled with the expectation of future rewards (Figure 5C); i.e., the early phasic activity was minimal when the next location was p_{25} and maximal when going for p_{100}. Because early phasic activity in pDAns was not caused by the previous ICSS but by learning (Figures 3 and 4), we assessed whether the scaling of VTA pDAn activity depended on factors such as prior delivery/omission of ICSS (R_t), the value of the current location (P_t), and the value of the next location (P_{t+1}). In a general linear model (GLM; Figure 5D) using these variables as predictors, VTA pDAns depended mostly on P_{t+1}, was to some extent influenced by R_t, but did not depend on P_t. Sorting VTA pDAns by prior reward and future location value (Figure 5E; Figures S7B and S7C) confirmed that both effects were independent, with the scaling of VTA pDAn activity as a function of (P_{t+1}), consistent with an expectation term.[35,43] pDAn response did not depend on the current location; e.g., there was no difference when the animal approached the p_{50} point from either p_{25} or p_{100} (Figure 5F), further confirming that animals' choices are guided by outcome expectation.

We thus further assessed whether DA cells also integrated the bonus value of uncertainty by computing the ratio between the VTA pDAn activity encoding of the most uncertain option (p_{50}) and of the most certain option (p_{100}). This p_{50}/p_{100} ratio correlated with φ (Figure 5G), which measures how much an animal values uncertain options. This suggests that VTA DAns integrate uncertainty with reward into a common currency[44] to promote the choice of uncertain options.[24]

PFC δ oscillation power did not scale with the expectation of reward uncertainty (Figure 5H). We analyzed PFC population firing (Figure 5I) during dwell time with a GLM similar to the one used for VTA pDANs (with P_{t+1}, P_t, and R_t), which was not significant against a constant model ($p = 0.277$). In contrast, a GLM using the expected reward uncertainty $(P_{t+1} \times [1 - P_{t+1}])$ of the next location showed that PFC firing depended on reward uncertainty but not on other predictors (Figure 5J). Accordingly, PFC population firing during dwell time was maximal when mice moved toward the location associated with the most uncertain (50%) reward probability (Figure 5J, S7D, and S7E). Furthermore, PFC population activity was enriched with the encoding of expected uncertainty (Figure S7F). Hence, contrary to PFC δ oscillations, which reflected prediction errors for preceding outcomes (Figure 4), population firing may signal the expected uncertainty of the upcoming outcome or at least a prediction of the most uncertain outcome. Finally, we did not find evidence of OFC θ oscillations (or firing) to encode expected reward or uncertainty (ANOVA $F_{(2)} = 0.3$, $p = 0.97$). Therefore, the internal representations of reward outcomes influencing animals' self-directions (i.e., expected reward and uncertainty) were represented in a complementary way by the VTA and PFC activities,

(H) Same as (C), left, for normalized PFC δ (3–6 Hz) power (a.u.) (mean δ power over 1–1.5 s after LE according to the probability of the goal: ANOVA, $F_{(2,60)} = 0.07$, $p = 0.93$).

(I) Same as (C) for PFC normalized population firing frequency (a.u.).

(J) Left: same as (D) for PFC normalized population firing frequency, with predictors: reward uncertainty of the goal (σ_{t+1}), P_t, and outcome delivered at current location (categorical variable, ICSS or omission). Right: mean PFC firing over 0.3–0.8 s after LE according to the probability of the goal: ANOVA, $F_{(2,147)} = 5.34$, $p = 0.006$. Data are presented as mean ± SEM.

***$p < 0.001$,**$p < 0.01$,*$p < 0.05$, n.s. $p > 0.05$.

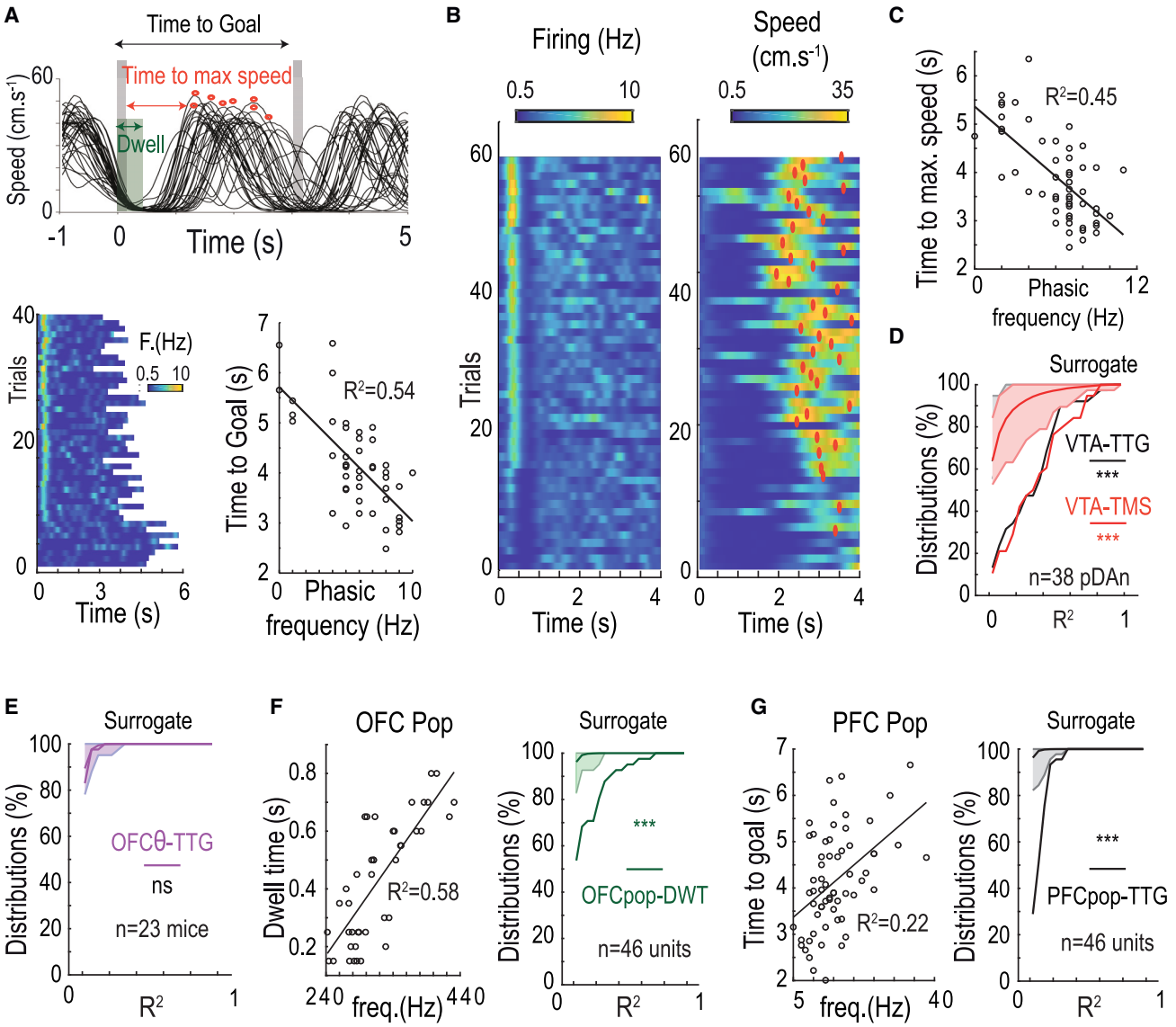


Figure 6. Distributed encoding of self-initiation, invigoration, and pacing by the OFC, VTA, and PFC

(A) Top: example of an instantaneous speed profile for one mouse in the D context. Gray boxes indicate ICSS durations, the green box the Dt, and red dots the maximal speed. Bottom: example of pDAn firing frequency in a D context session, with trials sorted from the smallest to the highest early phasic frequency (left) and relation between the time to goal and early phasic frequency (right), with each dot representing a trial.

(B) Example of firing frequency for another pDAn in the D context, with trials sorted from the smallest to the highest phasic activity frequency after trial initiation (left), and the instantaneous speed profile associated for each trial (right), with red dots indicating the maximal speed.

(C) Relation between the Ts within trials and the phasic frequency for the cell shown in (B).

(D) Distribution of correlation coefficients (R^2) between pDAn phasic frequency and time to goal (black) or Ts (red) in the D context for 38 cells. Data are presented as mean ± SEM across units. Surrogate data are generated by computing correlations with shuffled firing frequency and time to goal (or Ts). Kolmogorov-Smirnov test of data versus surrogates: $p < 10^{-3}$ for time to goal and Ts.

(E) Distribution of R^2 for correlations between OFC θ (7–14 Hz) power and Ts in the D context. Data are presented as mean ± SEM across subjects. Surrogate data are generated by computing correlation with shuffled θ power and Ts. Difference with surrogates: Kolmogorov-Smirnov test, $p = 0.62$.

(F) Left: example of relation between the Dt and the OFC population firing frequency in one mouse at the end of a D context session, with each dot representing a trial. Right: distribution of R^2 for correlations between OFC population firing frequency and Dt. Data are presented as mean ± SEM across units. Surrogate data are generated by computing correlation with shuffled firing frequency and Dt. Difference with surrogates: Kolmogorov-Smirnov test, $p < 10^{-3}$.

(G) Same as (F) for PFC population firing frequency and time to goal. Difference with surrogates: Kolmogorov-Smirnov test, $p < 10^{-3}$.

***$p < 0.001$,**$p < 0.01$,*$p < 0.05$, n.s. $p > 0.05$.

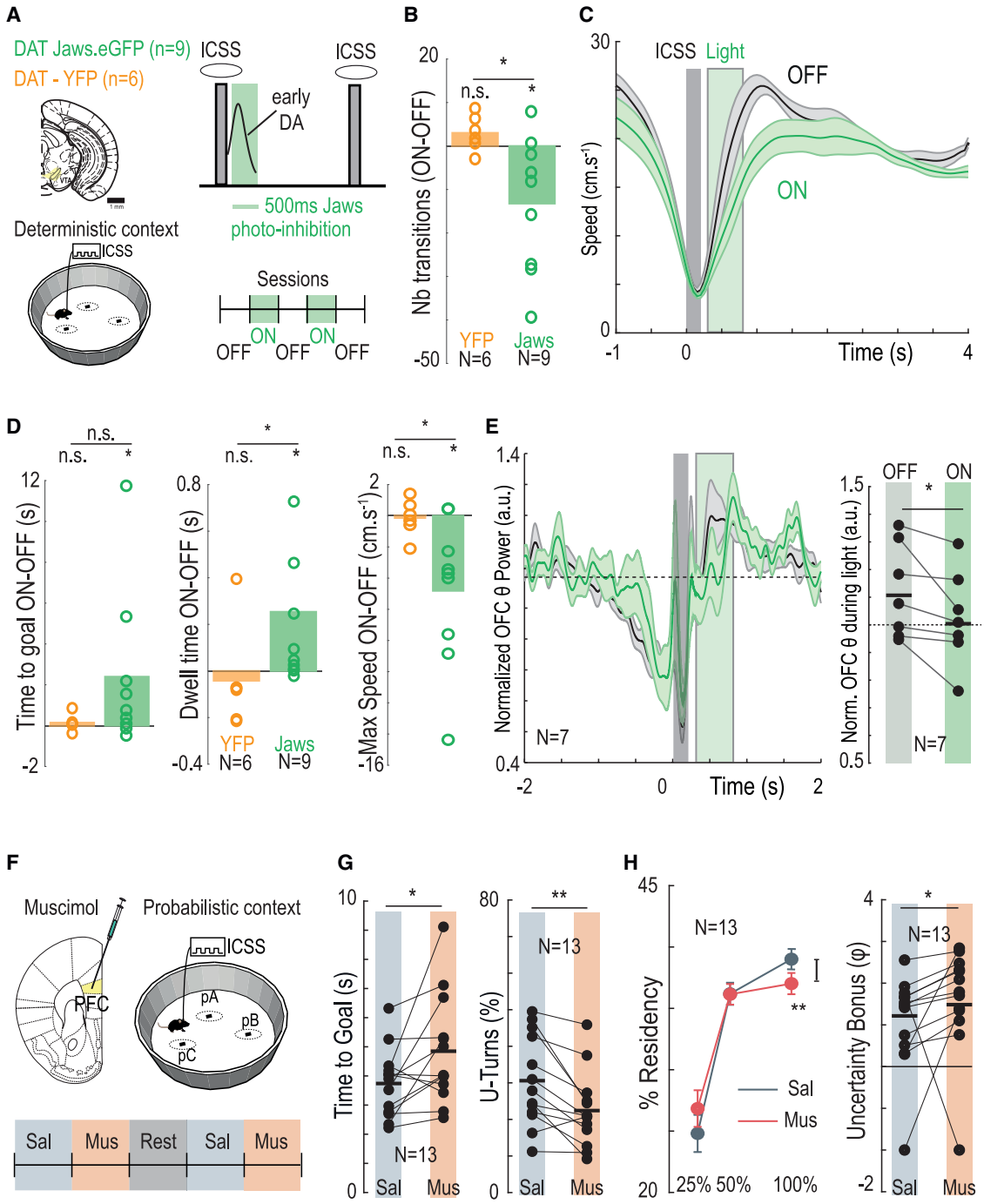


Figure 7. Behavior-dependent synergy or antagonism between DA VTA neurons and frontal cortices

(A) Manipulation experiment using the inhibitory opsin Jaws expressed unilaterally in the VTA of DATi^{CRE} mice using the CreLox strategy. Control animals were transduced with a YFP vector. Light was applied continuously for 500 ms, 100 ms after the end of the previous ICSS (STAR Methods), to suppress early phasic pDAn activity. Mice underwent, after the end of the D context, a succession of light OFF and ON sessions.

(B) Effect of light ON stimulation compared with OFF on the number of transitions (i.e., rewards obtained) (interaction between light and group conditions: $F_{(1,13)} = 6.06$, $p = 0.03$, paired Student's t test in the Jaws group: $T_{(8)} = 2.54$, $p = 0.03$, $\Delta = -13.2$). Horizontal bars represent means.

(C) Effect of light OFF or light ON stimulation on instantaneous speed profile. Data are presented as mean ± SEM across subjects. A gray box indicates the ICSS duration and a green box the light duration.

(D) Effect of light ON stimulations compared with OFF on the time to goal (left), Dt (center), and maximal speed (right). Time to goal: no interaction between light and group conditions: $F_{(1,13)} = 1.73$, $p = 0.17$, but paired two-sided Wilcoxon signed-rank test in the Jaws group: $W_{(8)} = 5$, $p = 0.04$, $\Delta = 2.40$ s; Dt: interaction between light and group conditions: $F_{(1,13)} = 5.57$, $p = 0.03$, paired two-sided Wilcoxon signed-rank test in the Jaws group, $W_{(8)} = 3$, $p = 0.0195$, $\Delta = 0.17$ s;

(legend continued on next page)

respectively, while the OFC (Figure 4) encoded a more general change in the task state.

Distributed correlates of self-initiation, invigoration, and pace of goal-directed actions

After examining the internal representations directing decisions, we then assessed the relation between circuit activity and the execution (initiation and invigoration) of self-paced actions. Each trial (Figure 1) is characterized by initiation of a movement toward the next location (estimated by the dwell time) and a strong acceleration (estimated by the time to maximal speed), followed by a deceleration at the next location (Figure 6A, top). Sorting the trials by ascending pDAn early phasic activity for each neuron revealed a negative correlation with the time to goal (Figure 6A, below) and the time to reach maximal speed (Figures 6B and 6C); greater pDAn phasic activity before self-initiation correlated with shorter time to goal because of a shorter time to reach the maximal speed. This was confirmed by the distribution of correlation coefficients (R^2) for all neurons (Figure 6D) and by use of surrogate data, which ruled out the possibility that these correlations were spurious (Figure 6D; STAR Methods).

In contrast, we found no significant correlation between cortical oscillation amplitude and successive behavioral timing, neither for OFC θ nor for PFC δ oscillations (Figures 6E and S7G). This might be due to the temporal order of neural oscillations and behavioral events; the increase in OFC θ and in PFC δ generally occurred after self-initiation (Figures 2 and 4) and thus may not be involved in this decision process. On the contrary, the increase in OFC population activity, which occurred early in the trial, correlated positively with the dwell time, suggesting involvement in self-initiation of the trial (Figure 6F). Finally, PFC population firing correlated with the time to goal but not with the dwell time or with the time to maximal speed (Figure 6G). This indicates that additional variability in the overall pace of the trial, which was not already due to earlier decisions (Figure 1), may be encoded in the PFC. Overall, OFC, VTA, and PFC firing activity synergistically encoded self-initiation, invigoration, and pace of the goal-directed actions, suggesting a sequential and distributed mechanism for self-paced decisions.

Synergy and antagonism between mesocortical structures in self-paced decisions under uncertainty

We next investigated the causal involvement of the VTA and PFC in selection (Figure 5) and execution of actions (Figure 6) by inactivating these structures during the task. To specifically manipulate VTA DAns, we expressed an inhibitory halorhodopsin variant (Jaws[45]) in DAT^{iCRE} mice using a Cre-dependent virus strategy (Figure 7A, left). We confirmed expression of the opsin in Jaws-transduced mice with immunochemistry and verified that 500-ms light pulses (520 nm) at 0.5 Hz reliably decreased the activity of VTA DAns using patch-clamp recordings (Figure S8A). Because the amplitude of the early phasic activity in VTA pDAns correlated with movement invigoration (Figure 6), we specifically tested the effect of optogenetic inactivation of VTA DAns at the time of the early phasic activity (500 ms of continuous light starting 100 ms after the previous ICSS; STAR Methods; Figure 7A, right). Optogenetic inhibition on each trial in the D context decreased the number of transitions (Figure 7B), which was due to an alteration of the speed profile that started during illumination and lasted after its termination (Figure 7C). In particular, photo-inhibition of VTA DAns delayed action initiation (increased dwell time; Figure 7D, center) and decreased action vigor (decreased maximum of mean speed; Figure 7D, right). None of these parameters were modified by light stimulation only in YFP-transduced DAT^{iCRE} mice (Figures 7B and 7D). Moreover, random photo-inhibition of VTA DAns in the home cage did not produce effects on speed (Figure S8B), suggesting that VTA DAn inhibition did not just slow speed regardless of motivation. Finally, we did not find any effect of sessions with VTA DAn photo-inhibition on subsequent sessions without photo-inhibition, suggesting that DA inhibition affected motivation directly rather than through learning (Figure S8C). These results causally implicate VTA DA cells in motivation for ongoing, self-generated movements, energizing movement as observed previously[20,41] but also promoting movement initiation[21] toward rewards.

VTA DA cell photo-inhibition also affected OFC θ oscillations (Figure 7E, left), with a decrease in θ oscillation power during (but not after) photostimulation (Figure 7E, right). We found no effect of VTA DA cell photo-inhibition on other OFC oscillation frequencies or on PFC oscillations. As OFC dynamics transitioned from increased population firing to θ oscillations at action initiation (Figures 2C and 7E), VTA DAn photo-inhibition may have directly or indirectly delayed θ oscillations, with OFC θ oscillations merely following the delayed action initiation. This latter interpretation is consistent with the peak in OFC θ power occurring at 0.62 s (and a dwell time of 0.42 s) without light and 0.83 s (with a dwell time of 0.67 s) under photo-inhibition.

maximal speed: interaction between light and group conditions: $F_{(1,13)} = 5.43$, $p = 0.03$, paired Student's t test, $T_{(8)} = 3.04$, $p = 0.016$, $\Delta = -4.81$ cm s^{-1}). Horizontal bars represent means.

(E) Left: effect of light ON stimulation compared with OFF on normalized OFC θ power (7–14 Hz) (paired Student's t test, $T_{(6)} = 2.63$, $p = 0.039$, $\Delta = -0.097$ a.u.). Data are presented as mean ± SEM across subjects. A gray box indicates ICSS duration and a green box light duration. Right: OFC θ power during light duration. Horizontal bars represent means.

(F) Schematic of the PFC inactivation experiment using a bilateral muscimol infusion at the end of the P context. Mice underwent a succession of sessions following saline or muscimol infusion.

(G) Effect of muscimol on the time to goal (left) and proportion of Uts (right) compared with saline (time to goal: paired two-sided Wilcoxon signed-rank test, $W_{(12)} = 12$, $p = 0.0171$, $\Delta = +1.11$ s; Uts: paired Student's t test, $T_{(12)} = 3.79$, $p = 0.0026$, $\Delta = -8.18\%$). Horizontal bars represent means across subject.

(H) Left: effect of muscimol on repartition between the three locations compared with saline (effect on the p_{100} choice: paired Student's t test, $T_{(12)} = 3.71$, $p = 0.003$, $\Delta = -1.99\%$). Data are presented as mean ± SEM across subjects. Right: effect of muscimol compared with saline on the fitted φ parameter obtained using the softmax based on the three parameters β, φ, and κ (STAR Methods) (paired two-sided Wilcoxon signed-rank test, $W_{(12)} = 13$, $p = 0.0215$, $\Delta = +0.54$). Horizontal bars represent means.

***$p < 0.001$,**$p < 0.01$,*$p < 0.05$, n.s. $p > 0.05$.

The absence of observable effects of VTA DA cell photo-inhibition on PFC dynamics is consistent with our electrophysiological results suggesting that the PFC is not implicated in self-initiation in the D context. In the P context, however, PFC δ oscillations and PFC firing frequency were correlated with time to goal and reward uncertainty. We thus inactivated the PFC bilaterally in the P context with local infusion of muscimol (Figures 7F and S8D). PFC inactivation in the P context increased the time to goal (Figure 7G, left) by increasing the dwell time and decreasing the maximal speed (Figure S8E, center and right). This resulted in a lower number of transitions (Figure S8E, left). These results confirm our electrophysiological data showing encoding of time to goal by PFC population firing (Figure 5) and suggest a role of the PFC in the overall pace of the action.

Surprisingly, muscimol in the PFC also decreased the percentage of U-turns (Figure 7G right). This was due to an altered choice repartition on the three locations (Figure 7H, left) with a decreased propensity to visit the location associated with the highest reward probability (i.e., 100%). We thus fitted the transition function of each mouse with the computational model (Figure S8F) in the saline and muscimol sessions (see STAR Methods and model U in Figure 5B). Under muscimol, animals behaved as if their φ was amplified (Figure 7H, right), with no changes in β or κ (Figure S8G). This increased φ did not correspond to an increase in residency on the 50% location but to a decrease in residency on the 100% location because the repartition on the rewarding locations arises from the sequence of binary choices in three Gs, which are not performed in equal proportions. Again, PFC inactivation with local muscimol extends our electrophysiological results; PFC population firing positively encoded reward uncertainty, and PFC inactivation increased the valuation of uncertainty, suggesting an inhibitory control of the PFC on uncertainty-seeking.

DISCUSSION

By using a task in which mice perform stereotyped trials from one rewarding location to another, we eliminated the requirement for an external stimulus to reiterate and time lock a specific behavior. Furthermore, the absence of a cue specifying the direction or initiation of each trial produced variability in action selection and execution, enabling correlation of behavior with neural activity. This allowed us to assess how self-generated actions arise from the contextual reorganization of mesocortical dynamics; a distributed sequence of firing and oscillations in the VTA, PFC and OFC jointly set the goal (where to go), self-initiated the trial, and determined the vigor and pace of the goal-directed action. This sequence was influenced by the reward context (D or P) and correlated with reward value and uncertainty, used to guide the animal's choices. Cortical oscillations and a distributed, transient increase in firing emerged during learning as a reorganization of existing dynamics, and all of these structures encoded prediction errors about the outcomes. Such a sequence, rather than being fixed, could incorporate the PFC or not, depending on the reward context, and the PFC could act in synergy with or antagonistically to the VTA in co-determining action selection or execution.

The MFB ICSS plays a critical role in our experiments. It is used as a reward surrogate, which allows us to focus our analysis on the emergence of internally generated activity based on learned, expected outcomes. ICSS has some advantage over natural rewards: it eliminates the need for food restriction and the associated satiation level, which could affect decisions, particularly under reward uncertainty.[46] In addition, because the time required to process the ICSS is very short compared with food consumption, mice engage in a rapid sequence of choices; many trials could be obtained in a relatively short time. Whether our observations are specific to the ICSS or generalizable to all rewards remains a question. ICSS may have resulted in saturating the value or learning functions, although model comparison favored an interpretation of animals' choices as uncertainty-seeking rather than saturating value function. The electrical stimulation may also have masked an increase in VTA DA cell activity by a random ICSS or early in training (see ICSS-responding and non-responding neurons in Figure S4). Other aspects of VTA DA cell activity at the time of the ICSS were, however, consistent with a classic RPE computation; pDAn activity is decreased at the time of the expected reward during unexpected omissions, whereas it is unchanged during expected ICSS, which may indicate an inhibition matching the antidromic activation. More importantly, ICSS consequences must be considered at two distinct time periods: during and after reward delivery (i.e., during the dwell time). Despite their temporal proximity, inherent to the task structure, these two time periods display different activities (i.e., reward activity and early phasic activity at the beginning of a new trial), reflecting distinct processes. Early phasic activity was observed after an expected and omitted ICSS but not after unexpected ICSS, demonstrating that it is not generated by the previous stimulation reward but, rather, emerges after learning. It is unlikely that the short dip in activity during omission leads to large rebound activity during the dwell time, as observed following a longer aversive stimulus.[47] Nevertheless, early phasic activity was not purely independent of prior ICSS or omission because VTA DA activity in the P context incorporated previous reward and expectation of the future outcome. Similarly, OFC and PFC activities were not solely the consequences of the previous ICSS but of learning.

It remains debated whether DA RPE at action initiation (corresponding to the early phasic activity in our task) reflects a learning signal[19,48] and/or constitutes a motivational command for the current trial.[18,21,41,49] During the dwell time, early phasic activity in VTA DAns scaled with the reward probability of the future location. By contrast, activity at the time of the expected reward decreased during reward omissions. These results comply with early phasic activity in VTA DA cells encoding an RPE. More importantly, early VTA DA phasic activity also correlated with movement vigor (estimated by the time to maximal speed and the time to goal). Furthermore, inhibition of this phasic activity reduced the time to goal by decreasing the maximal speed and delaying the action initiation but only in a rewarding context (i.e., not in the home cage). Although we cannot rule out that DA inhibition at trial initiation had a learning effect, the absence of cumulative effects on subsequent sessions favors the interpretation that VTA DA inhibition acutely affected the ongoing trial. Thus, in contrast to the view that

VTA DA activity is merely a passive reflection of action initiation,[19] we causally implicate RPE-like DA activity in the initiation and vigor of goal-directed actions.[18,21] We found no encoding of speed or acceleration in the PFC or OFC (which encoded the global pace and initiation time, respectively), suggesting that the effect of VTA DA on locomotion vigor may be mediated elsewhere, particularly in the nucleus accumbens.[2,17] Acutely inhibiting VTA DA cells delayed initiation so that VTA DA activity would facilitate a switch between states[21,49] (here, dwelling at a reward location) and goal-directed locomotion. The transition from increased spiking to θ oscillations at action initiation are in line with a state theory of the OFC,[31,50] where the OFC computes a "you are here" signal (within the task space) based on external (cues) and inferred information. Action initiation would constitute a change in the animal's state, associated with distinct OFC dynamics. This may also explain why OFC firing and oscillations react to reward and omission to signal changes between the different possible states of this task. However, in our task, action initiation is neither caused by an overt, transient stimulus cue nor by a hidden or inferred state (e.g., imposed delay[8]). On the contrary, OFC firing computed in advance the duration before state change (i.e., the dwell time), suggesting active involvement in controlling the change between task states (a "you wait here" signal) rather than a passive role in monitoring task states. Because OFC firing, but not pDAn firing, encoded the time of initiation in advance, while pDAn inhibition delayed initiation, this suggests a core OFC-VTA computation setting the self-initiation of a trial.

We did not see any correlation between OFC firing or θ oscillations and the expected reward or the reward uncertainty. This in line with a general role related to task space but at odds with accounts involving the OFC in economic value[13] or confidence (i.e., the inverse of uncertainty[51,52]). However, value or uncertainty can be confounded with arousal and salience, and causal involvement of OFC in economic choice is lacking.[53] By contrast, we found correlates of value and uncertainty in VTA DA cell firing and of uncertainty/error processing in PFC firing. Encoding of P rewards complies with RPE theories and has been described at length.[43,54,55] PFC δ oscillations have also been implicated in motivation.[56] However, expected uncertainty was not aggregated with expected value in PFC activity, contrary to observations in humans.[57,58] This might relate to task differences rather than species differences. Indeed, the PFC has been implicated in selecting the strategy.[15,31] In our task, we did not observe any increased firing or any PFC δ oscillations in the D context, suggesting that the PFC is mostly needed in the P context; i.e. for decisions under uncertainty or in the presence of prediction errors.[59,60] Because mice used uncertainty to guide their decisions in the P context, the PFC may compute decision-guiding heuristics that depend on uncertainty rather than uncertainty-modulated values per se.

Inhibiting the PFC with a local infusion of muscimol led to an increase in uncertainty-seeking, suggesting an inhibitory influence of the PFC on uncertainty-biased choices. Because DA generally has a positive influence on uncertainty-seeking,[24,61] this suggests antagonistic influence of the VTA and PFC on the motivation induced by reward uncertainty. This might reflect a difference between how model (or belief)-based and model-free control may treat uncertainty. In the P context, the uncertainty associated with the reward is known by the animals; i.e., it is a form of expected uncertainty.[62] In the simplest form of curiosity, expected uncertainty or variability may have a positive motivational influence in the form of a bonus added to the expected value by DA cells to promote exploration of unpredictable options.[24,63] By contrast, in deliberative strategies putatively implicating the PFC, the known, expected uncertainty may be treated as uninformative noise that has to be discarded from the decision strategy.[62] Hence, expected uncertainty may be incorporated by VTA DAns into value to promote model-free exploration and encoded by the PFC to favor model-based exploration, thus opposing model-free uncertainty-seeking.

Strikingly, the VTA and PFC had opposite influence on uncertainty-related choices but synergistic influence on the pace of actions. Suppressing the early phasic activity in pDAns and inhibiting the PFC similarly decreased the number of transitions. More work is needed to dissect whether PFC-DA interactions constitute sequential, recurrent, or independent computations[29,64,65] in ongoing, self-generated decisions. However, we suggest that this circuit presents a flexible organization as described for invertebrate circuits;[66] depending on the reward context, the PFC may flexibly integrate the VTA-OFC core circuit in charge for action initiation, adding uncertainty-based computations to the distributed sequence. In the same vein, the emergence of these activities with learning followed a reorganization of existing circuit dynamics. Indeed, with the completion of learning, a combination of bursts by DA cells at each trial, with an increase in the number of trials, could have resulted in an overall increase in pDAn firing frequency, but we did not observe any change. Hence, early phasic activity in DAns did not rely on additional spikes but, rather, on a dynamic re-organization toward time-locked bursting activity. This increase in early DA activity also correlated with locking of θ oscillations in the OFC. This suggests that the characteristics of the VTA-OFC-PFC (i.e., distributed but distinct contributions to learning and decisions) may rely on aligning distributed dynamics at relevant timings rather than on increases in activities of separate modules, each computing a decision variable. While this alignment of mesocortical dynamics is usually forced by a stimulus, we show here that it reorganizes even without a stimulus, resulting in self-generation of goal-directed actions.

Limitations of the study

Using electrical stimulation as a reinforcer has significant advantages, but the stimulation artifact blurs what is happening at the electrophysiological level when the animal receives a reward. Another limitation is with the treatment of stimulation reward; even if what we observe seems generalizable to all rewards, some specific mechanisms may be at work. Finally, electrophysiological recordings are only obtained in a limited number of brain regions, which cannot be considered the only regions involved in the behavioral processes studied, nor can these regions be thought of as being dedicated exclusively to performance in the task. The role of these regions and the mechanisms described must therefore be thought of as part of a global framework and not as an exclusive framework.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - AAV production
  - Intracranial self-stimulation electrode and recording electrode implantation
  - Virus injections
  - Polyelectrodes
  - Bipolar electrodes
  - Immunochemistry
  - Intracranial self-stimulation (ICSS) bandit task
  - Electrophysiological recordings
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Behavioral data analysis
  - Electrophysiological data analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.112523.

ACKNOWLEDGMENTS

We are grateful to the animal facilities (IBPS) and the Camille Robert and Paris Vision Institute AAV production facility for virus production and purification. This work was supported by the CNRS, INSERM, the Foundation for Medical Research (FRM; Equipe FRM DEQ2013326488 to P.F.), French National Cancer Institute grants TABAC-16-022 and TABAC-19-020 (to P.F.), and French state funds managed by the ANR (ANR-16-CE16-0020 and ANR-17-CE16-0016 to P.F.) and Memolife Labex Starting package (to P.F.).

AUTHOR CONTRIBUTIONS

Conceptualization, J.N. and P.F.; methodology, J.N., A.M., S.D., S.T., and P.F.; formal analysis, J.N., E.B., S.D., S.T., and P.F.; investigation, E.B., S.D., S.T., C.P.-S., M.C., J.N., T.A.Y., and E.V.; resources, L.T., writing – original draft, J.N., E.B., and P.F.; writing – review & editing, J.N., A.M., E.B., and P.F.; supervision, J.N. and P.F.; funding acquisition, P.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 5, 2022
Revised: January 28, 2023
Accepted: May 2, 2023
Published: May 17, 2023

REFERENCES

1. Passingham, R.E., Bengtsson, S.L., and Lau, H.C. (2010). Medial frontal cortex: from self-generated action to reflection on one's own performance. Trends Cognit. Sci. *14*, 16–21. https://doi.org/10.1016/j.tics.2009.11.001.
2. Klaus, A., Silva, J.A. da, and Costa, R.M. (2019). What, if, and when to move: basal ganglia circuits and self-paced action initiation. Annu. Rev. Neurosci. *42*, 1–25. https://doi.org/10.1146/annurev-neuro-072116-031033.
3. Cohen, J.D., McClure, S.M., and Yu, A.J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos. Trans. R. Soc. Lond. B Biol. Sci. *362*, 933–942. https://doi.org/10.1098/rstb.2007.2098.
4. Costa, R.M. (2011). A selectionist account of de novo action learning. Curr. Opin. Neurobiol. *21*, 579–586. https://doi.org/10.1016/j.conb.2011.05.004.
5. Schultz, W. (2007). Multiple dopamine functions at different time courses. Annu. Rev. Neurosci. *30*, 259–288. https://doi.org/10.1146/annurev.neuro.28.061604.135722.
6. Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593.
7. Gershman, S.J., and Uchida, N. (2019). Believing in dopamine. Nat. Rev. Neurosci. *20*, 703–714. https://doi.org/10.1038/s41583-019-0220-7.
8. 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.e6. https://doi.org/10.1016/j.neuron.2018.03.036.
9. Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., and Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat. Neurosci. *14*, 1590–1597. https://doi.org/10.1038/nn.2957.
10. Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. Philos. Trans. R. Soc. Lond. B Biol. Sci. *362*, 1585–1599. https://doi.org/10.1098/rstb.2007.2054.
11. Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. Annu. Rev. Neurosci. *34*, 333–359. https://doi.org/10.1146/annurev-neuro-061010-113648.
12. Hunt, L.T., and Hayden, B.Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. Nat. Rev. Neurosci. *18*, 172–182. https://doi.org/10.1038/nrn.2017.7.
13. Cai, X., and Padoa-Schioppa, C. (2014). Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. Neuron *81*, 1140–1151. https://doi.org/10.1016/j.neuron.2014.01.008.
14. Cisek, P. (2012). Making decisions through a distributed consensus. Curr. Opin. Neurobiol. *22*, 927–936. https://doi.org/10.1016/j.conb.2012.05.007.
15. Miller, E.K., and Wallis, J.D. (2013). Fundamental Neuroscience, Fourth Edition (Vii Behav Cognitive Neurosci), pp. 1069–1089. https://doi.org/10.1016/b978-0-12-385870-2.00050-0.
16. Berke, J.D. (2018). What does dopamine mean? Nat. Neurosci. *21*, 787–793. https://doi.org/10.1038/s41593-018-0152-y.
17. Salamone, J.D., and Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. Neuron *76*, 470–485. https://doi.org/10.1016/j.neuron.2012.10.021.
18. Fischbach-Weiss, S., Reese, R.M., and Janak, P.H. (2018). Inhibiting mesolimbic dopamine neurons reduces the initiation and maintenance of instrumental responding. Neuroscience *372*, 306–315. https://doi.org/10.1016/j.neuroscience.2017.12.003.
19. Coddington, L.T., and Dudman, J.T. (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat. Neurosci. *21*, 1563–1573. https://doi.org/10.1038/s41593-018-0245-7.
20. Howe, M.W., and Dombeck, D.A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. Nature *535*, 505–510. https://doi.org/10.1038/nature18942.
21. Silva, J.A. da, Tecuapetla, F., Paixão, V., and Costa, R.M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. Nature *554*, 1–21. https://doi.org/10.1038/nature25457.

22. Balleine, B.W. (2019). The meaning of behavior: discriminating reflex and volition in the brain. Neuron *104*, 47–62. https://doi.org/10.1016/j.neuron.2019.09.024.

23. Belkaid, M., Bousseyrol, E., Cuttoli, R.D., Dongelmans, M., Duranté, E.K., Yahia, T.A., Didienne, S., Hanesse, B., Come, M., Mourot, A., et al. (2020). Mice adaptively generate choice variability in a deterministic task. Communications Biology *3*, 1–9. https://doi.org/10.1038/s42003-020-0759-x.

24. Naudé, J., Tolu, S., Dongelmans, M., Torquet, N., Valverde, S., Rodriguez, G., Pons, S., Maskos, U., Mourot, A., Marti, F., and Faure, P. (2016). Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. Nat. Neurosci. *19*, 471–478. https://doi.org/10.1038/nn.4223.

25. Dongelmans, M., Durand-de Cuttoli, R., Nguyen, C., Come, M., Duranté, E.K., Lemoine, D., Brito, R., Ahmed Yahia, T., Mondoloni, S., Didienne, S., et al. (2021). Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration in mice. Nat. Commun. *12*, 6945. https://doi.org/10.1038/s41467-021-27268-7.

26. Carlezon, W.A., and Chartoff, E.H. (2007). Intracranial self-stimulation (ICSS) in rodents to study the neurobiology of motivation. Nat. Protoc. *2*, 2987–2995. https://doi.org/10.1038/nprot.2007.441.

27. Roesch, M.R., Calu, D.J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat. Neurosci. *10*, 1615–1624. https://doi.org/10.1038/nn2013.

28. Takahashi, Y.K., Batchelor, H.M., Liu, B., Khanna, A., Morales, M., and Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. Neuron *95*, 1395–1405.e3. https://doi.org/10.1016/j.neuron.2017.08.025.

29. Jo, Y.S., and Mizumori, S.J.Y. (2016). Prefrontal regulation of neuronal activity in the ventral tegmental area. Cerebr. Cortex *26*, 4057–4068. https://doi.org/10.1093/cercor/bhv215.

30. Lak, A., Okun, M., Moss, M.M., Gurnani, H., Farrell, K., Wells, M.J., Reddy, C.B., Kepecs, A., Harris, K.D., and Carandini, M. (2020). Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. Neuron *105*, 700–711.e6. https://doi.org/10.1016/j.neuron.2019.11.018.

31. Sharpe, M.J., Stalnaker, T., Schuck, N.W., Killcross, S., Schoenbaum, G., and Niv, Y. (2019). An integrated model of action selection: distinct modes of cortical control of striatal decision making. Annu. Rev. Psychol. *70*, 53–76. https://doi.org/10.1146/annurev-psych-010418-102824.

32. Grace, A.A., and Bunney, B.S. (1984). The control of firing pattern in nigral dopamine neurons: burst firing. J. Neurosci. *4*, 2877–2890.

33. Faure, P., Tolu, S., Valverde, S., and Naudé, J. (2014). Role of nicotinic acetylcholine receptors in regulating dopamine neuron activity. Neuroscience *282*, 86–100. https://doi.org/10.1016/j.neuroscience.2014.05.040.

34. Recorla, R.A., and Wagner, A. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Classical Conditioning II: Current Research and Theory, A.H. Black and W.F. Prokasy, eds. (Appleton-Century-Crofts), pp. 64–99.

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

36. Dayan, P., and Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. Cognit. Affect Behav. Neurosci. *8*, 429–453. https://doi.org/10.3758/cabn.8.4.429.

37. Takahashi, Y.K., Roesch, M.R., Stalnaker, T.A., Haney, R.Z., Calu, D.J., Taylor, A.R., Burke, K.A., and Schoenbaum, G. (2009). The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron *62*, 269–280. https://doi.org/10.1016/j.neuron.2009.03.005.

38. Yun, M., Kawai, T., Nejime, M., Yamada, H., and Matsumoto, M. (2020). Signal dynamics of midbrain dopamine neurons during economic decision-making in monkeys. Sci. Adv. *6*, eaba4962. https://doi.org/10.1126/sciadv.aba4962.

39. Murray, E.A., and Rudebeck, P.H. (2018). Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. Nat. Rev. Neurosci. *19*, 404–417. https://doi.org/10.1038/s41583-018-0013-4.

40. Roesch, M.R., Calu, D.J., Esber, G.R., and Schoenbaum, G. (2010). All that glitters . Dissociating attention and outcome expectancy from prediction errors signals. J. Neurophysiol. *104*, 587–595. https://doi.org/10.1152/jn.00173.2010.

41. Wassum, K.M., Ostlund, S.B., and Maidment, N.T. (2012). Phasic mesolimbic dopamine signaling precedes and predicts performance of a self-initiated action sequence task. Biol. Psychiatr. *71*, 846–854. https://doi.org/10.1016/j.biopsych.2011.12.019.

42. Glimcher, P.W. (2011). Quantification of Behavior Sackler Colloquium: understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc. Natl. Acad. Sci. USA *108*, 15647–15654. https://doi.org/10.1073/pnas.1014269108.

43. Tobler, P.N., Fiorillo, C.D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. Science (New York, N.Y.) *307*, 1642–1645. https://doi.org/10.1126/science.1105370.

44. Fiorillo, C.D. (2011). Transient activation of midbrain dopamine neurons by reward risk. Neuroscience *197*, 162–171. https://doi.org/10.1016/j.neuroscience.2011.09.037.

45. Chuong, A.S., Miri, M.L., Busskamp, V., Matthews, G.A.C., Acker, L.C., Sørensen, A.T., Young, A., Klapoetke, N.C., Henninger, M.A., Kodandaramaiah, S.B., et al. (2014). Noninvasive optical inhibition with a red-shifted microbial rhodopsin. Nat. Neurosci. *17*, 1123–1129.

46. Schuck-Paim, C., Pompilio, L., and Kacelnik, A. (2004). State-dependent decisions cause apparent violations of rationality in animal choice. PLoS Biol. *2*, e402. https://doi.org/10.1371/journal.pbio.0020402.

47. Robinson, J.E., Coughlin, G.M., Hori, A.M., Cho, J.R., Mackey, E.D., Turan, Z., Patriarchi, T., Tian, L., and Gradinaru, V. (2019). Optical dopamine monitoring with dLight1 reveals mesolimbic phenotypes in a mouse model of neurofibromatosis type 1. Elife *8*, e48983. https://doi.org/10.7554/elife.48983.

48. Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nat. Neurosci. *16*, 966–973. https://doi.org/10.1038/nn.3413.

49. Syed, E.C.J., Grima, L.L., Magill, P.J., Bogacz, R., Brown, P., and Walton, M.E. (2016). Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat. Neurosci. *19*, 34–36. https://doi.org/10.1038/nn.4187.

50. Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. Neuron *81*, 267–279. https://doi.org/10.1016/j.neuron.2013.11.005.

51. 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. https://doi.org/10.1038/nature07200.

52. O'Neill, M., and Schultz, W. (2010). Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. Neuron *68*, 789–800. https://doi.org/10.1016/j.neuron.2010.09.031.

53. Stalnaker, T.A., Cooch, N.K., and Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. Nat. Neurosci. *18*, 620–627. https://doi.org/10.1038/nn.3982.

54. Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. Nat. Neurosci. *19*, 479–486. https://doi.org/10.1038/nn.4239.

55. Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. Nature *577*, 671–675. https://doi.org/10.1038/s41586-019-1924-6.

56. Knyazev, G.G. (2012). EEG delta oscillations as a correlate of basic homeostatic and motivational processes. Neurosci. Biobehav. Rev. *36*, 677–695. https://doi.org/10.1016/j.neubiorev.2011.10.002.

57. Lebreton, M., Jorge, S., Michel, V., Thirion, B., and Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. Neuron *64*, 431–439. https://doi.org/10.1016/j.neuron.2009.09.040.

58. Tobler, P.N., Christopoulos, G.I., O'Doherty, J.P., Dolan, R.J., and Schultz, W. (2009). Risk-dependent reward value signal in human prefrontal cortex. Proc. Natl. Acad. Sci. USA *106*, 7185–7190. https://doi.org/10.1073/pnas.0809599106.

59. Rushworth, M.F.S., and Behrens, T.E.J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. Nat. Neurosci. *11*, 389–397. https://doi.org/10.1038/nn2066.

60. Bach, D.R., and Dolan, R.J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. Nat. Rev. Neurosci. *13*, 572–586.

61. Onge, J.R.S., and Floresco, S.B. (2010). Prefrontal cortical contribution to risk-based decision making. Cerebr. Cortex *20*, 1816–1828. https://doi.org/10.1093/cercor/bhp250.

62. Yu, A.J., and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron *46*, 681–692. https://doi.org/10.1016/j.neuron.2005.04.026.

63. Anselme, P., Robinson, M.J.F., and Berridge, K.C. (2013). Reward uncertainty enhances incentive salience attribution as sign-tracking. Behav. Brain Res. *238*, 53–61. https://doi.org/10.1016/j.bbr.2012.10.006.

64. Ellwood, I.T., Patel, T., Wadia, V., Lee, A.T., Liptak, A.T., Bender, K.J., and Sohal, V.S. (2017). Tonic or phasic stimulation of dopaminergic projections to prefrontal cortex causes mice to maintain or deviate from previously learned behavioral strategies. J. Neurosci. *37*, 8315–8329. https://doi.org/10.1523/jneurosci.1221-17.2017.

65. Park, J., and Moghaddam, B. (2017). Risk of punishment influences discrete and coordinated encoding of reward-guided actions by prefrontal cortex and VTA neurons. Elife *6*, e30056. https://doi.org/10.7554/elife.30056.

66. Marder, E., and Bucher, D. (2007). Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. Annu. Rev. Physiol. *69*, 291–316. https://doi.org/10.1146/annurev.physiol.69.031905.161516.

67. Fobbs, W.C., and Mizumori, S.J.Y. (2014). Cost-benefit decision circuitry: proposed modulatory role for acetylcholine. Prog. Mol. Biol. Transl. Sci. *122*, 233–261. https://doi.org/10.1016/b978-0-12-420170-5.00009-x.

68. Daw, N.D. (2011). Trial-by-trial data analysis using computational models. In Decision Making, Affect, and Learning Decision Making, Affect, and Learning (Oxford University Press), pp. 3–38. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001.

69. Yu, Z., Guindani, M., Grieco, S.F., Chen, L., Holmes, T.C., and Xu, X. (2022). Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron *110*, 21–35. https://doi.org/10.1016/j.neuron.2021.10.030.

70. Le Van Quyen, M., and Bragin, A. (2007). Analysis of dynamic brain oscillations: methodological advances. Trends Neurosci. *30*, 365–373. https://doi.org/10.1016/j.tins.2007.05.006.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-tyrosine Hydroxylase produced in mouse	Sigma-Aldrich	Cat# T1299, RRID:AB_477560
Anti-GFP produced in chicken	Aveslabs	Cat# GFP-1020, RRID:AB_10000240
Anti-rabbit Cy2-conjugated produced in donkey	Jackson ImmunoResearch	Cat# 711-225-152, RRID:AB_2340612
Anti-mouse Cy3-conjugated produced in donkey	Jackson ImmunoResearch	Cat# 715-165-150, RRID:AB_2340813
Anti-chicken Alexa 488-conjugated	Jackson ImmunoResearch	Cat# 703-545-155, RRID:AB_2340375
Bacterial and virus strains		
AAV5.EF1α.DIO.Jaws.eGFP	This paper: Provided by Institut de la vision, Paris France	Virus (AAV)
AAV5.EF1α.DIO.YFP	This paper Provided by Institut de la vision, Paris France	Virus (AAV)
Chemicals, peptides, and recombinant proteins		
NaCl	Sigma-Aldrich	S7653
KCl	Sigma-Aldrich	P9333
NaH2PO4	Sigma-Aldrich	S8282
MgCl2	Sigma-Aldrich	M2670
CaCl2	Sigma-Aldrich	233506
NaHCO3	Sigma-Aldrich	S6297
Sucrose	Sigma-Aldrich	S0389
Glucose	Sigma-Aldrich	49159
Kynurenic Acid	Sigma-Aldrich	K3375
Albumin, from bovine serum	Sigma-Aldrich	A4503
KGlu	Sigma-Aldrich	P1847
HEPES	Sigma-Aldrich	H3375
EGTA	Sigma-Aldrich	E3889
ATP	Sigma-Aldrich	A9187
GTP	Sigma-Aldrich	G8877
Biocytin	Sigma-Aldrich	B4261
Nicotine tartrate	Sigma-Aldrich	N5260
Glucose	Sigma-Aldrich	G8270
DPBS 10x	Life Technologies	14200–067
Neurobiotin Tracer	Vector laboratories	SP-1120
Prolong Gold Antifade Reagent	Invitrogen	P36930
Chloral Hydrate	Sigma-Aldrich	302-17-0
Sodium Acetate	Sigma-Aldrich	57654611
Quinpirole	Tocris	55397
Eticlopride	Tocris	57266
Muscimol	Tocris	0289
Deposited data		
Raw and analyzed data	This paper	Table S1
Experimental models: Organisms/strains		
Mouse: C57Bl/6Rj	Janvier Laboratories, France	SC-C57J-M
Mouse: C57Bl/6Rj DAT^{iCRE}	Turiault et al., 2007 https://doi.org/10.1111/j.1742-4658.2007.05886.x	DATicre maintained on a C57BL6/J background

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R Project for Statistical Computing	http://www.r-project.org/	RRID:SCR_001905
MATLAB	Mathworks	RRID:SCR_001622
Adobe Illustrator 2020	Adobe	RRID:SCR_010279
Cheetah version 3.01 2.5.4	Neuralynx	Neuralynx acquisition
SpikeSort3D	Neuralynx	Neuralynx acquisition
Clampfit (pClamp suite)	Molecular Devices	RRID:SCR_011323
Labview	National Instruments	RRID:SCR_014325

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Philippe Faure (phfaure@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data Availability: All the data that support the findings of this study can be found in the Source Data file provided with the paper. Raw data are available from the corresponding authors.

Code Availability: All codes used to run the analysis are available from the authors upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Experiments were performed on adult C57Bl/6Rj DAT^{iCRE} and Wild-Type (Janvier Labs, France) mice. 64 male mice, from 8 to 16 weeks old, weighing 25–35 g, were used for all the experiments. They were kept in an animal facility where temperature (20 ± 2°C) and humidity were automatically monitored and a circadian light cycle of 12/12-h light-dark cycle was maintained. All experiments were performed in accordance with the recommendations for animal experiments issued by the European Commission directives 219/1990, 220/1990 and 2010/63, and approved by Sorbonne University.

Experimental units correspond to

- N = 12 WT mice were implanted with VTA electrodes, 23 WT mice were implanted with bilateral OFC/PFC electrodes, 16 DAT^{iCre} mice were injected with opsins/YFP vectors, 13 WT mice were implanted with PFC cannulas.
- N = 31 mice over the 35 (12 + 23) electrode-implanted mice, for which we have all the training sessions in all contexts, were used for behavioral analyses and computational model (Figures 1, 3, and 5). 4 mice were discarded because of corrupted video detection on some sessions.
- Baseline behaviors for DAT^{iCre} and cannulas-implanted groups were analyzed separately, to replicate the findings on WT (Figure 7).
- For bipolar electrodes, EFP signals were analyzed by mouse (N = 23 or less if signal was lost for a session). When both bilateral signals were valid, they were averaged to have 1 OFC or PFC data per animal/session (Figures 2, 3, and 4). Population firings in these structures were analyzed separately, yielding N = 46, or less if there was no spiking activity for a session (Figures 2, 4, 5, and 6).
- For VTA multi-electrodes, the experimental unit was the neuron (N = 136 throughout learning contexts) because averaging was not always possible, since on some learning sessions there was no neurons for a given animal (Figures 2, 3, 4, 5, and 6).

METHOD DETAILS

AAV production

AAV vectors were produced as previously described using the cotransfection method and purified by iodixanol gradient ultracentrifugation.[51] AAV vector stocks were tittered by quantitative PCR (qPCR)[52] using SYBR Green (Thermo Fischer Scientific).

Intracranial self-stimulation electrode and recording electrode implantation

Mice were anesthetized with a gas mixture of oxygen (1 L/min) and 1–3% of isoflurane (Piramal Healthcare, UK), then placed into a stereotaxic frame (Kopf Instruments, CA, USA). After the administration of a local anesthetic (Lurocain, 0.1 mL at 0.67 mg/kg), a median incision revealed the skull which was drilled at the level of the Median Forebrain Bundle (MFB), the OFC, the PFC or the VTA. Dental cement (SuperBond, Sun Medical) was used to fix the implant to the skull. A bipolar stimulating electrode for ICSS was then implanted unilaterally (randomized) in the brain (stereotaxic coordinates from bregma according to mouse after Paxinos atlas: AP -1.4 mm, ML ±1.2 mm, DV -4.8 mm from the brain). Bipolar recording electrodes were implanted in the lateral OFC (AP +2.6 mm, ML ±1.5 mm, DV -1.7 mm from the brain) and the medial PFC (AP +1.65 mm, ML ±0.5 mm, DV -1.8 mm from the brain). Multi-electrodes were implanted in the VTA (AP -3.15 to −3.25 mm, ML ±0.5 mm, DV -4.1 to 4.25 mm from the brain). After stitching and administration of a dermal antiseptic, mice were then placed back in their home-cage and had, at least, 5 days to recover from surgery. An analgesic, buprenorphine solution at 0.015 mg/L (0.1 mL/10 g), was delivered after the surgery and if necessary, the following recovering days. The efficacy of electrical stimulation was verified through the rate of acquisition during the deterministic context (see behavioral STAR Methods).

Virus injections

DAT^{iCRE} mice were anesthetized (Isoflurane 1–3%) and were injected unilaterally (randomized left/right side and ipsi/contralateral side) in the VTA (1 μL, coordinates from bregma: AP -3.15 to −3.25 mm; ML ±0.5 mm; DV -4.55 mm from the skull) with an adeno-associated virus (AAV5.EF1α.DIO.Jaws.eGFP $1.16e^{13}$ ng/μL or AAV5.EF1α.DIO.YFP $6.89e^{13}$ or $9.10e^{13}$ ng/μL). A double-floxed inverse open reading frame (DIO) allowed to restrain the expression of Jaws (red-shifted cruxhalorhodopsin) to VTA dopaminergic neurons.

Polyelectrodes

Hand-made multi-electrodes (2 bundles of 8 electrodes) were obtained by twisting eight polyimide-insulated 17 μm Nickel-Chrome wires. The use of eight channels relatively close together allows for a better discrimination of the different neurons. Before implantation and recording, the multi-electrodes were cut at suitable length and plated using a Platinium-PEG solution to lower their impedance to 150–400 KOhms and improve the signal-to-noise ratio. The free ends of the multi-electrodes were connected to the holes of EIB-18 (electrode interface board, Neuralynx) and fixed with pins. We manufactured a microdrive system (home-made 3D conception and printing) consisting of a main body, on which is mounted the EIB, and a driving screw, with a sliding part design to contain the two multi-electrodes. This microdrive allowed moving through the VTA in order to sample neuronal populations.

Bipolar electrodes

Hand-made bipolar electrodes were obtained by twisting two Teflon-insulated (60 μm) Stainless Steel wires. Two configurations were used. For the first one, the tips of the bipolar electrodes were cut so that they are spaced of less than 0.5 mm apart. For the second one, the reference tip was wound around the recording one, at a distance of less than 0.5 mm from the recording endpoint. These electrodes are designed so the two tips are oriented perpendicular to the dipoles formed by cortical pyramidal neurons. The first configuration was used for OFC recording electrodes, and the second one for PFC recording electrodes. IntraCranial Self-Stimulation (ICSS) electrodes were made as the second configuration with an 80 μm Stainless Steel wire. Bipolar electrodes were connected to the EIB during the surgery, by fixing the free ends with pins.

Immunochemistry

After euthanasia, brains were rapidly removed and fixed in 4% paraformaldehyde (PFA). After a period of at least three days of fixation at 4°C, serial 60-μm sections were cut with a vibratome (Leica). Immunostaining experiments were performed as follows: VTA brain sections were incubated for 1 h at 4°C in a blocking solution of phosphate-buffered saline (PBS) containing 3% bovine serum albumin (BSA, Sigma; A4503) (vol/vol) and 0.2% Triton X-100 (vol/vol), and then incubated overnight at 4°C with a mouse anti-tyrosine hydroxylase antibody (anti-TH, Sigma, T1299) at 1:500 dilution, in PBS containing 1.5% BSA and 0.2% Triton X-100. The following day, sections were rinsed with PBS, and then incubated for 3 h at 22–25°C with Cy3-conjugated anti-mouse and secondary antibodies (Jackson ImmunoResearch, 715-165-150) at 1:500 in a solution of 1.5% BSA in PBS, respectively. After three rinses in PBS, slices were wet-mounted using Prolong Gold Antifade Reagent (Invitrogen, P36930). Microscopy was carried out with a fluorescent microscope, and images captured using a camera and analyzed with ImageJ. In the case of optogenetic experiments on DAT^{iCRE} mice, identification of the transfected neurons by immunohistofluorescence was performed as described above, with the addition of 1:500 Chicken-anti-GFP primary IgG (ab13970, Abcam) in the solution. A Goat-anti-chicken Alexa Fluor 488 (1:500, Life Technologies) was then used as secondary IgG. Neurons labeled for TH in the VTA allowed to confirm their neurochemical phenotype, and those labeled for GFP to confirm the transfection success.

Intracranial self-stimulation (ICSS) bandit task

Behavioral set up

The ICSS bandit task took place in a circular open field with a diameter of 68 cm. Three explicit square-shaped marks (1 × 1 cm) were placed in the open field, forming an equilateral triangle (side = 35 cm). Entry in the circular zones (diameter = 6 cm) around each mark

was associated with the delivery of a rewarding ICSS stimulation. Experiments were performed using a video camera, connected to a video-tracking system, out of sight of the experimenter. A LabVIEW (National Instruments) application precisely tracked and recorded the animal's position with a camera (20 frames/s). When a mouse was detected in one of the circular rewarding zones, an electrical stimulator received a TTL signal from the software application and generated a 200 ms-train of 0.5-ms biphasic square waves pulsed at 100 Hz (20 pulses per train). ICSS intensity was adjusted, within a range of 20–200 μA, during training (see training contexts) and then kept constant, so that mice would achieve between 50 and 150 visits per session (5min duration) for two successive sessions, and then kept constant for all the experiment. The constant motivational level insured by ICSS alleviated the need for a stimulus to repeat the behavior. Mice with insufficient scores in the PS and DS (<40 visits despite increasing the intensity to a maximum of 200 μA) were excluded.

Baseline *behavior*

Prior to the ICSS bandit task, three control sessions were performed. First, spontaneous neuronal activity was recorded in the mice home-cages for 10 min. Second, neuronal activity was recorded while random ICSS were delivered to the mice in its home-cage, to assess the direct effect of the stimulation onto neuronal activity. Third, behavioral and neuronal activity were recorded for 30 min, while the mice were exploring the open-field for the first time ("habituation", without the presence of the three rewarding locations).

Training *context*

The training consisted of two context s: the deterministic context (D) and the probabilistic context (P), consisting of 10 daily sessions of 5 min for the DS and 10 min for the PS. In the DS, all zones were associated with an ICSS delivery (p = 100%). However, two consecutive rewards could not be delivered on the same location, which motivates mice to alternate between locations. In the PS, the zones were associated with three different probabilities (p = 25%, p = 50%, p = 100%) to obtain an ICSS stimulation. The probabilities' locations were pseudo-randomly assigned per mouse. Animals successively make the task in DS and then in PS.

Data acquisition per experimental group

For optogenetics experiments, the DAT^{iCRE} mice (n = 16) completed the training, followed by a schedule of 4 days of paired sessions with photo-stimulation (ON) alternated with days without photostimulation (OFF). The averages of the ON and OFF days were compared in a paired manner.

Optogenetics *experiments*

For optogenetic experiments on freely moving mice, an optical fiber (200 μm core, NA = 0.39, Thor Labs) coupled to a ferule (1.25 mm) was implanted just above the VTA ipsilateral to the viral injection (coordinates from bregma: AP -3.1 mm, ML ±0.5 mm, DV 4.4 mm), and fixed to the skull with dental cement (SuperBond, Sun Medical). The behavioral task began at least 4 weeks after virus injection to allow the transgene to be expressed in the target dopamine cells. An ultra-high-power LED (520 nm, Prizmatix) coupled to a patch cord (500 μm core, NA = 0.5, Prizmatix) was used for optical stimulation (output intensity of 10 mW). Optical stimulation during the behavioral experiment was continuously delivered for 500 ms, starting 100 ms after animal's detection in a location. The ON and OFF schedule (OFF-ON-OFF-ON-OFF) was following the last week of deterministic training. The optical stimulation cable was plugged onto the ferrule during 5 experimental sessions to prepare the animals and control for latent experimental effects.

Intracranial *injections* of muscimol

A solution of muscimol (TOCRIS) (0.5μg/μL) was infused in the PFC over 20–30 min before the beginning of the ICSS bandit task experiment. The bilateral infusion of 0.4μL was performed at a rate of 0.2μL/min using a double injector (Univentor). Before each experiment session, a double injection cannula (2.5 mm, 0.5 mm projection) was inserted into the implanted bilateral cannula guide (length below pedestal 2.5 mm). The injection cannula was connected to a multi-syringe pump (Univentor) that allowed saline or muscimol injection. The saline and muscimol schedule (saline-muscimol-rest-saline-muscimol) was following the last week of probabilistic training. The injection system was plugged onto the cannula guide before 5 experimental sessions to prepare the animals and control for latent experimental effects.

Electrophysiological recordings

All extracellular potentials recordings were performed using a digital acquisition system (Digital Lynx SX; Neuralynx) together with the Cheetah software. Broadband signals from each wire were filtered between 0.1 and 9000 Hz and recorded continuously at 32 kHz.

Multi-unit activity record*ings*

To extract spike timing, signals were band-pass filtered between 600 and 6000 Hz and sorted offline. Spike clustering was cross-validated by using both SpikeSort3D (Neuralynx) and custom-written MATLAB (The Mathworks) routines. The electrophysiological characteristics of VTA neurons were analyzed in the active cells encountered by systematically moving down the multi-electrodes.

Local-field potential recordings

To extract low-frequency variations of extracellular potential, signals were low-pass-filtered below 300 Hz.

Population firing

To extract spike timing of the neuronal population, signals were band-pass filtered between 600 and 6000 Hz and sorted offline. Because population firing originates from bipolar electrodes with only one recording wire, no clustering could be considered.

ICSS artifacts

Electrical stimulation of the MFB induced artifacts during the 200ms train of pulses. These artifacts could be clustered in the multi-unit recordings and thus VTA DA activity during the ICSS could be recorded, with the potential caveat that the spikes concomitant with the 0.5ms pulses were discarded. Population firing was not clustered, hence we did not consider population activity during the 200ms

ICSS duration. This period is marked with a gray mask, e.g. in Figures 2 and 4, and the apparent dip in activity is only caused by the removal of activity during the 200ms. Finally, the (filtered) effect of the ICSS artifact could not be removed from EFP recordings without altering the wavelet transforms. Hence, we did not remove any signal, but did not analyze the wavelet transform during the 200ms and 100ms before and after the train duration (400 ms total) to avoid border effects. We thus only analyzed the ICSS period for VTA DAn activity (in any condition), but not for OFC/PFC EFP or population firing.

Identification of DA cells

Extracellular identification of putative DA neurons (pDAn) was based on their location as well as on a set of unique electrophysiological properties that characterize these cells *in vivo*: 1) a typical triphasic action potential with a marked negative deflection; 2) a characteristic long duration (>2.0 ms) action potential; 3) an action potential width from start to negative trough >1.1 ms; 4) a slow firing rate (<12 Hz) with an irregular single spiking pattern and occasional short, slow phasic activity. Putative GABA neurons were characterized by a characteristic short duration of action potential from start to negative trough (<1.0 ms), and a high firing rate (>12 Hz). D2 receptors (D2R) pharmacology was also used for confirming the DA neurons identification: after a baseline period (5 min) and a saline (10 min) injection, quinpirole (1 mg/kg, D2R antagonist) was injected (30 min recording), followed by an eticlopride (D2R agonist) injection (1 mg/kg, 10 min recording). Since most DA, but not GABA neurons, express inhibitory D2 auto-receptors, neurons were considered as pDA neurons if quinpirole induced at least 30% decrease in their firing rate, while eticlopride restored firing above the baseline. Nevertheless, as continuous D2 pharmacology could have affected both baseline DA neurons firing and decision-making,[67] we allowed the mice to recover two days after this experiment. We thus performed pharmacological confirmation (1) when first encountering a putative DA neuron in a given mouse or (2) at the end of the week if at least one putative neuron was present during the behavioral experiment. Neurons were considered as pDAn only if they responded to the pharmacology, or if they presented electrophysiological characteristics defined above and were recorded between two positive pharmacological experiments.

Ex vivo patch-clamp recordings

To verify the functional expression of Jaws, an AAV5.EF1a.DIO.Jaws.eGFP virus was injected into the VTA of 7 to 9-week-old male DATiCRE mice. After 4 weeks, coronal midbrain sections (250 μm) were prepared as already described in 25. Briefly, slices were transferred to a recording chamber continuously perfused at 2 mL/min with oxygenated aCSF, which contained (in mM): 125 NaCl, 2.5 KCl, 1.25 NaH2PO4, 2 CaCl2, 1 MgCl2, 26 NaHCO3, 15 Sucrose, and 10 Glucose (pH 7.2, 325 mOsm). Whole-cell recordings were performed with a patch-clamp amplifier (Axoclamp 200B, Molecular Devices) connected to a Digidata (1550 Low Noise acquisition system, Molecular Devices). Patch pipettes (4–8 MΩ) were pulled from thin wall borosilicate glass (G150TF-3, Warner Instruments) using a micropipette puller (P-87, Sutter Instruments, Novato, CA) and filled with a KGlu-based intra-pipette solution containing (in mM): 116 K-gluconate, 10-20 HEPES, 0.5 EGTA, 6 KCl, 2 NaCl, 4 ATP, 0.3 GTP, and 2 mg/mL biocytin (pH adjusted to 7.2). Optical stimulation was applied through the microscope with a 520 nm LED (CoolLED). Recordings were made in the voltage-clamp (-60 mV, continuous photostimulation 1 s) or current-clamp mode (train of ten stimulations, 500 ms, 1 Hz). Signals were low-pass filtered (Bessel, 2 kHz) and collected at 10 kHz using the data acquisition software pClamp 10.5 (Molecular Devices). All electrophysiological recordings were extracted using Clampfit (Molecular Devices) and analyzed with R.

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavioral data analysis

Behavioral *measures*

For all groups of mice, the trajectory was smoothed using a triangular filter allowing the determination of speed profile, which corresponds to instantaneous speed as a function of time, and time of maximal speed within a trial. The following measures were analyzed in the DS and compared in the PS, as well as in the DS for the OFF vs. ON Jaws experiment, or in the PS for the Sal vs. Mus experiment: i) number of visits, ii) time-to-goal, iii) choice repartition (proportion of visits p_{25}, p_{50} and p_{100}), iv) percentage of directional changes (n^{th} visit = n^{th} visit+2). Furthermore, the ICSS bandit task can be seen as a Markovian decision process. Every transition between zones can be considered as a binary choice between two probabilities, since the occupied zone cannot be reinforced twice in a row. The sequence of choices per session is summarized by the proportional result of the sum of three specific binary choices (or gambles, i.e., total visits zone 1/total visits zone 1 + 2). The three gambles (G) were named after the point on which the mouse is positioned at the time of the choice: G_{25} = 100% vs. 50%, G_{100} = 50% vs. 25% and G_{50} = 100% vs. 25%.

Locomotor activity toward the rewarding locations was measured in terms of time-to-goal, dwell time and time to maximal speed. Time-to-goal measures the duration between one location and the next one. The speed profile corresponds to the instantaneous speed as a function of time (20 frames per s). The dwell time is defined as the duration between the end of the 200 ms period (corresponding to the eventual ICSS duration) in the last rewarding location and the moment when the animal's speed is greater than 10 cm s^{-1}. The time to maximal speed is the time at which the speed profile attains its maximal value. We compared general linear regression models (GLM) of the time-to-goal with increasing number of explanatory variables (with Bayesian information criterion). Best explanatory variables were whether the animal performed a U-turn, the dwell time, and the time to maximal speed (minus the dwell time to remove its additive influence). We regularized the GLM for correlated terms using ridge regression, insuring that each predictive variable exerted an uncorrelated effect on the time-to-goal. We finally checked that each parameter had a significant influence ($p < 0.05$) on the time-to-goal for each animal.

Modeling

The location choice in these gambles reflects the balance between exploitative (choosing the most valuable option) and exploratory (choosing the least valuable option) choices. With a softmax based decision-making model fitted in the laboratory, we computed three parameters: the value sensitivity or inverse temperature (the power to discriminate between values in a binary choice), the uncertainty bonus (the preference for expected uncertainty, considering the reward variance of every option in a binary choice) and the motor cost to do a directional change (a decrease in the location value if it requires to go back to the previous location). Decision-making models determined the probability P_i of choosing the next state I, as a function (the ''choice rule'') of a ''decision variable''. Because mice could not return to the same rewarding location, they had to choose between the two remaining ones. Accordingly, we modeled decisions between two alternatives labeled A and B and used a softmax choice rule defined by $P_{A\ =}1/(1+e^{-\beta(vA-vB)})$ where β is an inverse temperature parameter reflecting the sensitivity of choice to the difference between decision variables and V_i the value of an option. The value V of an option is modeled as the expected (average) reward + expected uncertainty + U-turn cost.[16,30] This compound value is then nested in the softmax choice rule, given a 6*3 matrix that described the probability of a choice between A, B and C (the three locations) depending on the two previous choices. As an example, in the probability to choose (A, B, C) after performing the sequence BA, the value is given by $(0,\ p_b+\varphi p_b*(1-p_b)-\kappa,\ p_c+\varphi*p_c*(1-p_c))$ while after the sequence CA the value is given by $(0,p_b+\varphi*p_b*(1-p_b),p_c+\varphi*p_c*(1-p_c)-\kappa)$ (same for AB, CB and AC, BC). The free parameters of the model were fitted by maximizing the data likelihood. Given a sequence of choice $c = c_{1..T}$, data likelihood is the product of their probability (given by previous equation).[68] We derived Bayesian Information Criterion from the likelihood and used it to compare the full model with simpler ones, i.e. a softmax model in which choices only depend on expected value (φ and $\kappa = 0$) and a softmax model in which choices depend on expected value and motor cost ($\varphi = 0$).The winning model in the probabilistic setting included: i) a value sensitivity parameter (β) measuring the trade-off between exploitative choices and random decisions, ii) a reward uncertainty bonus measuring how much animals value uncertain options and iii) a motor cost (κ) measuring the negative value of performing a U-turn,[23,25] as mice rather than U-turns (Figure 1), in both contexts. As an alternative to uncertainty-seeking, we evaluated a saturating value function, in which value saturates with reward probability: the value of the largest reward probability ($p = 1$) was set at 1-s, with s a saturation parameter comprised between 0 (no saturation) and 0.5 (in which case the $p = 0.5$ and $p = 1$ options have the same value). We also checked that simpler models (null model of random choice, null model with a motor cost, epsilon-greedy with constant exploration) did not provide a better fit. We used the *fmincon* function in MATLAB to perform the fits, with the constraints that $\beta \in\ ]0,10]$, $\varphi \in\ ]-1,5]$ and $\kappa \in\ ]0,5]$.

Statistical analyses

All statistical analyses were computed using MATLAB and Python with custom programs. Results were plotted as a mean ± s.e.m. The total number (n) of observations in each group and the statistics used are indicated in figure legends. Classical comparisons between means were performed using parametric tests (Student's T test, or ANOVA for comparing more than two groups) when parameters followed a normal distribution (Shapiro test $p > 0.05$), and non-parametric tests (here, Wilcoxon or Mann-Whitney) when the distribution was skewed. Multiple comparisons were Bonferroni corrected. Probability distributions were compared using the Kolmogorov–Smirnov (KS) test, and proportions were evaluated using a chi-squared test (χ^2). Unless stated, statistical analyses of unit (spiking) activity implicitly assumed fixed effects across subjects rather than random effects.[69] We checked that the results on probability encoding by VTA DA cells, and uncertainty encoding by PFC population activities (Figure 5), as well as encoding of locomotion variables (Figure 6) were robust after taking into account inter-subject variability, by using three kinds of surrogate analyses (see Figure S7). For results in Figure 5, we generated an ensemble ($n = 10000$) of ''resampled animals'' datasets by resampling subjects from the original dataset with replacement (such than in any given ''resampled'' dataset, units from each animal may appear multiple times or not at all), and checked that the ANOVA was generally significant, indicating that the statistics are robust and does not depend exclusively on few neurons from one single individual. We also generated an ensemble ($n = 10000$) of ''shuffled encoding'' datasets, by shuffling the firing activities encoding the (25%, 50%, 100%) reward probability for every animal (so that for all units of a given animal, neuronal activities are shuffled in the same way, e.g. 50% encoding becomes for example 25% encoding in the surrogate). We then checked that ANOVA was generally not significant for these permutations, indicating that the results are driven by reward encoding rather than by inter-individual differences (i.e dependencies between units recorded in the same animal level do not produce high Type I error rates). For locomotion encoding (Figure 6), as the correlated variable (e.g. time to goal) was continuous, we shuffled the firing activity for each neuron at the level of trials, computed the R2 distribution for a given repetition ($n = 10000$), and obtained the confidence interval of the distribution by a jackknife method to assess whether the experimental R2 distribution significantly differed from the surrogate distributions.

Electrophysiological data analysis

Firing *analysis*

Spontaneous DA cell firing was analyzed with respect to the average firing rate and the percentage of spikes within bursts (%SWB, number of spikes within bursts, divided by total number of spikes). Bursts were identified as discrete events consisting of a sequence of spikes such that: their onset is defined by two consecutive spikes within an interval <80 ms and they terminated with an interval >160 ms. Phasic activity is defined as spikes falling into bursts, while tonic activity comprises spikes outside bursts. Peri-event time histograms (PETH) for normalized activity were constructed based on 1 ms-bins rasters, convolved with a Gaussian kernel (100 ms, using 50ms or 200ms did not change the results), divided by the neuron basal firing rate (to compare DA neurons with firing rates from 1 to 10Hz). Normalized PETH were sorted according to the *preceding* event (reward or omission) in Figures 2 and 4,

and to the probability of reward associated with the *next* location in Figure 5. Phasic activity from these PETH was defined as the firing rate during a 500ms time window (usually 300-ms-800ms after last location entry unless stated in the Results). We checked that the results did not depend on the exact time window by systematically shifting the beginning (100ms–500ms) and duration (300ms–800ms) of the time windows by 50ms bins. Encoding of reward uncertainty by PFC multi-unit activity was also assessed through an enrichment analysis: we determined for which reward probability of target location the PFC population activity was the highest, intermediate and lowest. PFC phasic activity was considered to encode uncertainty if it was highest for 50%, intermediate for 25%, and lowest for 100% probability. The proportion of PFC activity encoding uncertainty was compared to expected proportion (there are 6 possible orders when sorting activities related to 3 events, giving 16.7% as expected proportion).

Wavelet analysis

Because extracellular field potentials (EFP) are non-stationary signals, they are transformed offline using a Morlet wavelet transform (center frequency = 0.6 and bandwidth = 1). This process is defined as the convolution product between the EFP signal and dilated forms of wavelets normalized to 1.[70] EFP signal was expressed in *Z* score units in Figure 2. For each channel, the *Z* score normalization used the mean and the standard deviation from the 2s period preceding the location entry (LE). In Figures 4, 5, and 7, EFP signal was also band-pass filtered in the θ (7–14 Hz) or δ (3–6 Hz) frequency band and normalized for each channel with the mean power in each frequency band. The cross-spectra (cross-correlograms between OFC and PFC power spectra) in Figure S2 were computed for brain regions of the same hemisphere and per animal. The wavelet coherence (normalized spectral covariance) between the EFP from the OFC and the one from the PFC was computed by smoothing the product of the two wavelet transforms over time (window for time smoothing = 0.2s) and over scale (pseudo-frequency) steps (window for scale smoothing = 2 Hz).