nature communications

Article

https://doi.org/10.1038/s41467-026-74783-6

Neural and computational correlates of strategic aborting and long-run policy optimization in the dorsolateral prefrontal cortex

Received: 28 November 2024

Accepted: 8 June 2026

Published online: 26 June 2026

Check for updates

Jean-Paul Noel[1,2,7] ✉, Ruiyi Zhang[3,4,5,6,7], Xaq Pitkow[3,4] & Dora E. Angelaki[5,6]

Real-world choices often require balancing short- and long-term goals. We reasoned that seemingly suboptimal single-trial decisions may reflect strategic planning over longer timescales. We demonstrate that male macaques freely navigating in virtual reality strategically aborted offers, forgoing immediate rewards to maximize session-long returns. This behavior was highly individual-specific, suggesting that macaques account for their own long-run performance. Reinforcement-learning models suggest that this strategy is supported by modular actor-critic networks in which a policy module optimizes long-term value while also incorporating state-action values for rapid policy adjustment. These models predict that policy changes for matched offers should emerge at offer presentation, even when aborts occur much later. Consistent with this prediction, units and population dynamics in dorsolateral prefrontal cortex (dlPFC), but not parietal area 7a or dorsomedial superior temporal area (MSTd), encoded upcoming reward-optimizing aborts at offer onset. These findings cast dlPFC as a specialized policy module within closed-loop behaviors.

In dynamic environments animals continuously select actions to maximize a utility function[1]. Namely, we select actions in an attempt to balance the acquisition of reward, now or in the foreseeable future, with the expenditure of resources[2] (e.g.,energy). To enable adaptive behavior, these policies driving action selection ought to be sensitive to changes in the value of available options, our uncertainty about the state of the world, and/or our ability to alter the environment. How the brain instantiates policies—e.g., deciding what actions are worth an effort —is not fully understood. In particular, we do not fully understand how strategic, long-term planning (e.g., beyond a single trial, or offer) is updated and maintained within dynamic, closed action-perception loops.

Significant progress toward answering this question has come from reinforcement learning[3] (RL) and the development of artificial neural network agents. In this setting, we can construct various network architectures, define utility functions, and examine learned policies. This approach, when applied to neuroscience, first led to the conjecture that dopamine-driven plasticity in the striatum translates experienced action-reward associations into optimized behavioral policies[4]. More recently, it has been suggested that a second neural RL algorithm exists, a meta-reinforcement learning system that is bootstrapped yet independent from the dopaminergic striatum[5,6]. This latter meta-RL network is hypothesized to (1) be reflected in neural activity (as opposed to synaptic weights) of the pre-frontal cortex, (2)

[1]Department of Neuroscience, University of Minnesota, Minneapolis, USA. [2]Minnesota Robotics Institute, College of Science and Engineering, University of Minnesota, Minneapolis, USA. [3]Neuroscience Institute, Carnegie Mellon University, Pennsylvania, USA. [4]Machine Learning Department, Carnegie Mellon University, Pennsylvania, USA. [5]Center for Neural Science, New York University, New York, USA. [6]Tandon School of Engineering, New York University, New York, USA. [7]These authors contributed equally: Jean-Paul Noel, Ruiyi Zhang. ✉e-mail: noelx071@umn.edu

be more model-based than the dopaminergic striatum[7] (but see[8]), and (3) more readily allow for generalization[9] (see[5] for further detail).

Prior work has largely supported claims casting the prefrontal cortex as implementing components of an RL algorithm. For instance, this area (or sub-divisions within) reflect(s) subjective value[10], decision confidence[11,12], the effort required to harvest a reward[13,14], and/or task strategy selection[15,16]. These seminal studies, however, have important limitations. First, most (primate) studies artificially dissociated periods of action, perception, and reward. Second, they typically only allow for small and discrete state-spaces (e.g., two-alternative forced choice).

Third, on a trial-by-trial basis offers are typically independent, not allowing for a naturalistic trade-off between short- and long-term returns. Indeed, in most traditional work, a correct choice now is rewarded, and an incorrect choice is not. But the lack of a reward now does not suppose an increased or accelerated potential for reward later. In existent exceptions, where animals can decide to opt-out of uncertain outcomes, this often involve trading an immediate potential high-value reward for an assured, but also immediate, low-value outcome. Fourth, *when* opting out is a possibility is almost always exogenously controlled by experimenters rather than endogenously controlled by the agents themselves. In other words, many tasks classically used to probe macaque prefrontal cortex neurophysiology do not unequivocally need model-based RL, and do not operate within naturalistic continuous action-perception loops. In fact, these tasks are not always modeled as an RL problem, and there is often no evidence that they require among other, generalization[5], long-term planning[17,18], and the forgoing of immediate reward for a better long-term return.

In an effort to study real-world computations, our group has developed a closed-loop task wherein macaques use a joystick controlling their linear and angular velocity to continuously path integrate in a virtual environment and stop at the location of transiently presented (300 ms) targets—akin to catching a briefly flashing firefly[19,20] (Fig. 1A). This task is executed in continuous state and action spaces, requires an evolving belief (i.e., posterior probabilities) over partially observable states, and is learned via sparse rewards (i.e., trials may last many seconds and animals are only rewarded at the end of a sequence of actions). Importantly, within this naturalistic context, animals generalize with no further training to variants of the task requiring novel sensorimotor mappings, inference, decision-making, and foraging[21]. That is, within this task, animals demonstrate the use of a model-based approach and zero- or one-shot generalization[21]. Further, we have previously developed actor-critic RL networks instantiating artificial agents performing this "firefly task." We found that modular architectures − separating the computation of belief from action (in the actor), and/or belief from value (in the critic) − help agents generalize quickly within this task[22] (also see[23] for a similar finding).

Here, we build on this approach, both experimentally and computationally, to examine (1) if and how macaques employ long-term strategic plans within this closed-loop task, (2) how they trade-off short- vs. long-term potential returns, (3) what algorithms or RL architectures they may use for implementing long-term policies, and (4) what neural codes may subserve these session-long policies.

Results

Macaques reason about their own performance

Three adult male macaques (Macaca mulatta; 9.8–13.1 kg) used a two-dimensional joystick to control their virtual linear (maximum = 200 cm/s) and angular (maximum = 90 deg/s) velocity to path integrate and stop at the location of transiently visible targets (72 sessions and ~100,000 trials total). Each trial began with a brief (300 ms) presentation of a circular target ("firefly") positioned 100–400 cm in depth and within ±45° azimuth (independent uniform distributions). Aside from the transient targets, the virtual environment consisted solely of a textured ground plane composed of randomly oriented and positioned triangles (lifetime = 250 ms; see Fig. 1A). These transient ground-plane elements provided depth perspective and optic-flow cues when the animal was in motion, enabling estimation of self-motion velocity and its integration into an evolving belief of self-position. Given that the ground-plane triangular elements are transient and re-positioned and oriented each time they are drawn, they do not provide landmark information. Each trial terminated when the animal either ceased movement or when the trial timed out after 7 s. A juice reward was delivered if the final position fell within 65 cm of the target. There was no statistical correlation between trials (e.g., target locations or repeat of unrewarded trials). In most trials (~ 88%), the animals were stationary during target presentation, though this was not required; in the remaining ~12% of trials, animals had a non-zero linear velocity at target onset (i.e., they were moving during the inter-trial interval).

In prior work we have described macaques' mean performance[20,21] and their neural correlates[24–26] in successfully navigating toward and stopping at the location of briefly presented targets ("fireflies") in virtual reality. The animals are rewarded on 61.5% of trials (mean across sessions), and the mean correlation between target and response location is typically $r^2 > \sim 0.75$ (usually larger for radial than angular distances and varying across animals; see[24]). Here, instead, we examine an equally important (and still numerous, 38.5%) set of trials: "failure modes" within this task. These are the trials that are arguably the most informative regarding the animals' strategies. We particularly focus on examining long-term policies (which may be sub-optimal on a shorter time-frame), and changes in this policy given matched experimental offers.

Examining the radial distance animals traveled as a function of the presented target distance (uniform distribution, 100-400 cm in radial distance, see "Methods") revealed four categories of unrewarded trials (see distribution for an example session in Fig. S1). The majority of these were attempted trials: the animals simply misestimated—typically undershot[20,21,24–26]—the location of targets (mean ± s.e.m across sessions; 52.6% ± 1.5% of unrewarded trials; Fig. 1B leftmost). In another two categories of unrewarded trials, the animals either never started navigating toward the target (13.5% ± 2.5% of unrewarded trials; Fig. 1B second column), or never stopped (17.4% ± 2.1% of unrewarded trials; Fig. 1B third column). In both these cases, the trial timed-out after 7 sec. Lastly, on 16.5% ± 2.2% of unrewarded trials, the animals started navigating toward the target (linear velocity threshold > 10 cm/s; minimum distance traveled > 5 cm) and then abruptly stopped, thus aborting the trial (maximum distance traveled <30% target distance; Fig. 1B rightmost). On average, animals aborted 114.3 trials per session (range of total trials per session; 449–2250 trials), and these aborts occurred on average 1.66 sec (± 0.09 seconds) after target presentation. The average attempted trial ended 2.88 sec (± 0.07 sec) after target presentation (aborted vs. attempted durations across trials, $t = 8.42$, $p < 0.001$). Aborted trials were largely non-directed with respect to the target: when attempted trials were temporally truncated to match aborted trial duration, the correlation between target angle and final movement direction was markedly reduced and was near zero for aborted trials (attempted average $r^2 = 0.33$; aborted $r^2 = 0.02$; $p < 10\text{-}6$), indicating that aborting reflects an early and deliberate policy decision rather than a partially executed, but misdirected, attempt.

The aborted trials appeared to be a deliberate task strategy, in contrast to the "never started" or "never ended" trials which are seemingly better described as disengaged states[27,28]. Namely, the transition probability between trial-types (Fig. 1C) showed that after an attempted trial (rewarded or not), animals most often attempted the next trial (89.3% of trials). Similarly, animals disproportionately repeated "never started" trials (69.2%) or "never stopped" trials (37.1%). In the "never stopped" trials it was common for animals to simply hold the joystick in a given position for 30+ seconds, resulting in circular motions. In contrast, after an *aborted* trial, the animals most often attempted the next trial (60.4%) rather than aborting again (22.1%,

https://doi.org/10.1038/s41467-026-74783-6

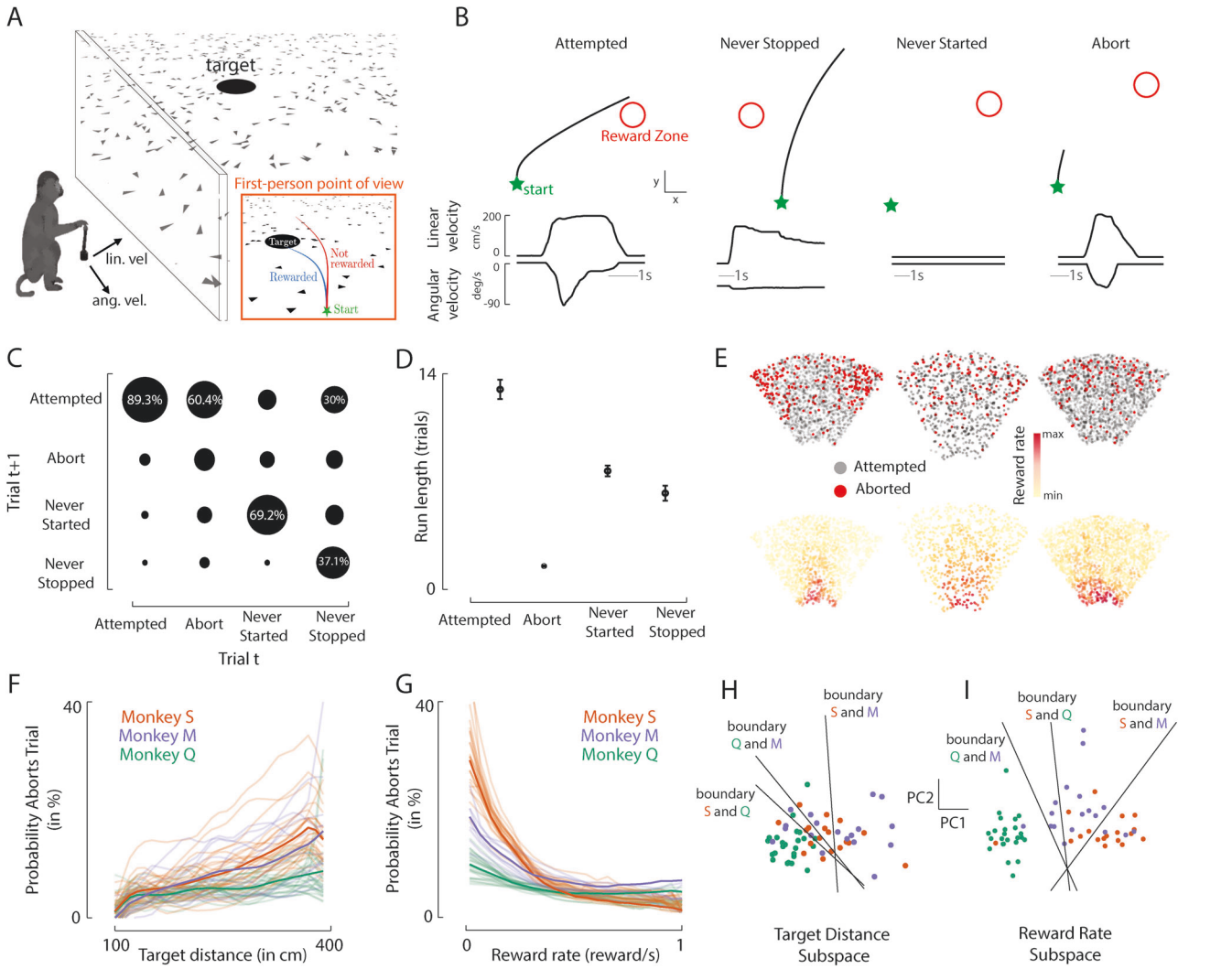


Fig. 1 | Macaques abort trials to maximize session-long reward rates. **A** Schematic of the task and reward structure. Animals use a joystick controlling their linear and angular velocity to path integrate to the location of a briefly presented (300 ms) target. Targets are presented within 100 to 400 cm of depth (radial distance), and −40° to +40° eccentricity (uniform distribution). Animals are rewarded for stopping within 65 cm of the target. Panel shows both an allocentric point of view (3 d) and as an inset (orange) a rendering of the first-person point of view of the macaque. Schematic is adapted from[21] with permission from the publisher. **B** Example trials showing stereotypical "failure modes": attempted yet unrewarded trials (leftmost), trials where the animal did not stop within the maximum trial time (7 sec, "never stopped", 2nd column), trials where the animal never started moving (3rd column), and trials where the animal accelerated and then abruptly aborted the trial (rightmost). Bottom of each column shows a time-series of linear (in cm/s) and angular (deg/s) velocity (gray inset shows 1 sec for each example trial). Top of each columns show a bird's-eye-view. **C** Transition probability matrix between trial types (trial *tr* along the x-axis and *tr* + 1 along the y-axis). Circle size is proportional to probability of transition. Data across all monkeys. **D** Run lengths of a particular trial type. Error bars are ± 1 standard error of the mean (s.e.m.). Data across all monkeys. **E** Three example sessions, one from each monkey (S, M, and Q). Top: Location of targets (overhead view) that were attempted (gray) or aborted (red). Bottom: same plot as above, while plotting the reward rate offered by a given targets (according to animal- and session-specific likelihood of obtaining reward for that target, and the time required to reach it). **F** Probability of aborting trial as a function of target distance (integrated sums to '1'). Each animal is plotted in a different color; opaque line shows the average for the specific animal, while translucent lines show individual sessions. **G** Same as **F**, plotted as a function of reward rate. Animals are less likely to abort trials that offer a high reward rate. **H** Data from **F** is plotted in its two leading principal components (colored as a function of animal) and linear discriminant analysis is applied to classify sessions by animal (black lines show bifurcation lines). **I** Same as (**H**) when projecting data from (**G**). Source data are provided as a Source Data file.

Chi-square test, $\chi^2 = 19.9$, $p = 2.9 \times 10^{-17}$). This same pattern can be observed by computing the number of consecutive trials of the same category (i.e., run-lengths, Fig. 1D). Animals on average attempted 13.9 trials in a row (± s.e.m = 0.6). Similarly, "never started" and "never ended" trials had relatively long run lengths, respectively of 7.6 (± 0.34) and 6.0 (± 0.47) trials. In contrast, the average run length of aborted trials was only 1.2 trials (± 0.05), suggesting that this behavior was intentional and not indicative of a longer-term disengagement state[27,28]. Analysis of eye movements also supported this conclusion: on aborted trials animals detected and saccade toward the target just as they did in attempted trials (Fig. S2A-D), while in "never started" and "never ended" trials the animals did not saccade during target presentation (Fig. S2E), and there was no meaningful correlation between the real and predicted gaze location under the hypothesis that animals look toward the target (Fig. S2F).

The experimentally-imposed variable that most clearly drove aborting behavior was the radial and angular distance of the presented target (Fig. 1E; three example sessions shown, one per animal). However, there was a large spatial overlap between trials animals chose to attempt (gray), and those they chose not to (red, Fig. 1E, see "salt-and-pepper" mixing of attempted and aborted trials). In other words, the animal's policy (e.g., attempt or not) could differ for a "matched offer"

(i.e., same location of target). In turn, we hypothesized that perhaps the factor determining whether animals aborted or not trials was the perceived afforded reward rate of a given trial. The afforded reward rate is not an experimentally-imposed variable: it depends on the animal- and session-specific probability of being rewarded for a given target presentation, and on how fast the animal completes the trial. Thus, if aborting behavior is driven by afforded reward rates, this behavior is animal- and session-specific.

We computed the reward rate afforded by targets in an animal- and session-specific manner (see "Methods"), and expressed the probability of aborting a trial as a function of either target distance (experimentally-imposed, Fig. 1F) or afforded reward rate (a variable determined in part by the macaque itself, Fig. 1G). Monkeys aborted more trials as the target distance (at presentation) grew, but the exact nature of this relationship varied across sessions (Fig. 1F, note large variability across sessions for a given animal), mimicking the variance observed on individual trials (Fig. 1E). Instead, when aborting probability is expressed as a function of reward rate (Fig. 1G), each session could be nicely attributed to a particular animal (Fig. 1G) and it becomes apparent that monkeys were preferentially aborting trials that offered a low reward for the time investment required. To quantitatively capture this pattern, we projected the probability of aborting trials as a function of distance (Fig. 1F) and reward rate (Fig. 1G) into their first two principal components (respectively accounting for 89.9% and 91.3% of the total variance). In these spaces, we performed linear discriminant analysis (LDA[29]) to classify held-out (10-fold cross-validation) sessions as belonging to one of three monkeys (chance $\sim$ 36.6% given the unequal number of sessions per animal). The classifier based on target distance achieved an accuracy of 70.0% ($\pm$ 0.8%; Fig. 1H), while the classifier based on reward rates achieved an accuracy of 89.6% ($\pm$ 0.7%; $p = 0.0004$; Fig. 1I; 28% better than the classifier based on target distance).

Together, these results suggest that macaques employed the strategy of deliberately aborting trials (forgoing reward on trial *tr*). On average, each abort "saved" the animal 1.22 sec. This behavior was animal- and session-specific, demonstrating their ability to take into account their own performance (in terms of probability correct and trial duration) and implement an adaptive long-term ($\sim$ session-long) plan.

Modular RL actors with access to specific values from critic strategically abort trials to maximize reward rate

So far, we have shown that when targets appear, monkeys deliberately evaluate the value of completing or aborting each trial. This behavior suggests a model-based decision process in which the animals' policy relies on a comparison between alternatives. To better understand the computational principles underlying this form of deliberation, and assess its functional benefits, we developed artificial agents that incorporate (1) action-selection not solely within trials, but also across them, and (2) an explicit value-comparison mechanism. We compare the RL agents developed here to standard RL agents, including those used in our previous work[22], which lack the long-term planning and explicit value-comparison mechanism.

We trained RL agents to navigate to the location of hidden targets instantiating different neural architectures, and examine if and how they abort offers (Fig. 2A). We formulate the "firefly task"[20,21,24–26] as a Partially Observable Markov Decision Process (POMDP[30,31]). Specifically, at each time step t, the state of the environment $\mathbf{s}_t$ includes the agent and target position, as well as the agent's velocity. The agent then makes an observation given the state, $\mathbf{o}_t \sim O(\mathbf{o}_t, |, \mathbf{s}_t)$, and takes an action $\mathbf{a}_t$ controlling its linear and angular velocity. The action updates the state to $\mathbf{s}_{t+1}$, according to the environment transition probabilities $T(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, i.e., the task dynamics. The agent receives reward r_t determined by the reward function $R(\mathbf{s}_t, \mathbf{a}_t)$, which is a positive scalar if the agent stops within the reward zone (65 cm, as for the macaques), and 0 otherwise.

To generate behaviors, we adopt an actor-critic approach[3,32,33] in training artificial agents to perform this task (Fig. 2A). At each time t the actor aims to generate an action $\mathbf{a}_t$ that maximizes the state-action value, which represents the expected sum of discounted rewards from time t onward:

$$\mathbb{E}_{\mathbf{s}\prime, \mathbf{a}\prime}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | \mathbf{s}_t, \mathbf{a}_t\right] \quad (1)$$

where γ is the temporal discount factor and $\mathbf{s}\prime$ and $\mathbf{a}\prime$ denote future states and future actions. In this work, there is no terminal state after a trial concludes, and instead, the selection of actions considers not only the reward associated with stopping at the target location in the current trial, but also the potential rewards for all future potential targets. This differs from our[22] (and most of the field's) previous work, in which the state-action value function was bound to the current trial (i.e., there was no consideration of time-frames outside the current trial). A critic network (Fig. 2A, B) is used to approximate state-action values, denoted as Q_t, given $\mathbf{s}_t$ and $\mathbf{a}_t$. The critic's approximation is updated using the temporal-difference reward prediction error[34] (TD error) δ_t,

$$\delta_t = r_t + \gamma Q_{t+1} - Q_t \quad (2)$$

where Q_{t+1} is the next-step value obtained through bootstrapping (for details see "Methods" and *Eqs. 18, 19* in the *Supplementary Information*). In our task, observations $\mathbf{o}_t$ only contain partial information about the state. For instance, the target is only visible for a short duration (i.e., the 300 ms over which it is visible). Likewise, self-position of the agent has to be inferred from velocity observations (i.e., "path integration") and is not directly observable (i.e., there is no landmark information in this task). Hence, the critic network first needs to construct an internal state representation, or 'belief' b_t, about the current state (the actual state being a feature the agent does not have access to) in order to approximate belief-action values. This belief is computed via a recurrent neural network (RNN; Fig. 2B) that enables the temporal integration of evidence and retains this information in its activity via recurrence. The input, $\mathbf{i}_t$, for the RNN includes the noisy observation $\mathbf{o}_t$ (self-velocities and target position when the target is visible) and the previous action $\mathbf{a}_{t-1}$ as an efference copy. Next, the critic approximates Q_t given b_t from the RNN and an action $\mathbf{a}_t$. This latter computation is accomplished via a feedforward (i.e., memoryless) multi-layer perceptron (MLP; Fig. 2B, see[22] for the advantages of this modular architecture).

Note that the critic including the RNN and MLP is trained end-to-end through back-propagation with the objective of learning Q_t. Nevertheless, in Zhang et al[22]. we showed that even without an explicit objective to shape b_t, the learned belief is near-optimal, and the RNN and MLP are respectively specialized in computing belief (b_t, Fig. 2B, cyan) and value (Q_t, Fig. 2B), without intermixing the computations of these variables. Therefore, we refer to the RNN as the belief module and the MLP as the value module. This terminology reflects the computational roles of each module, rather than the variables they represent. In particular, although belief-related information is necessary within the value module because it takes b_t as input, its primary computational function is value estimation. The same convention applies to the actor network. An actor network is trained to select an action $\mathbf{a}_t$ that maximizes the value Q_t of the critic (Fig. 2C-E). The algorithm treats the actor and critic as separate networks, as they are optimized under different objectives (see "Methods"). Therefore, in addition to generating actions, the actor also needs a belief state derived from raw sensory inputs. This belief can either be computed by

https://doi.org/10.1038/s41467-026-74783-6

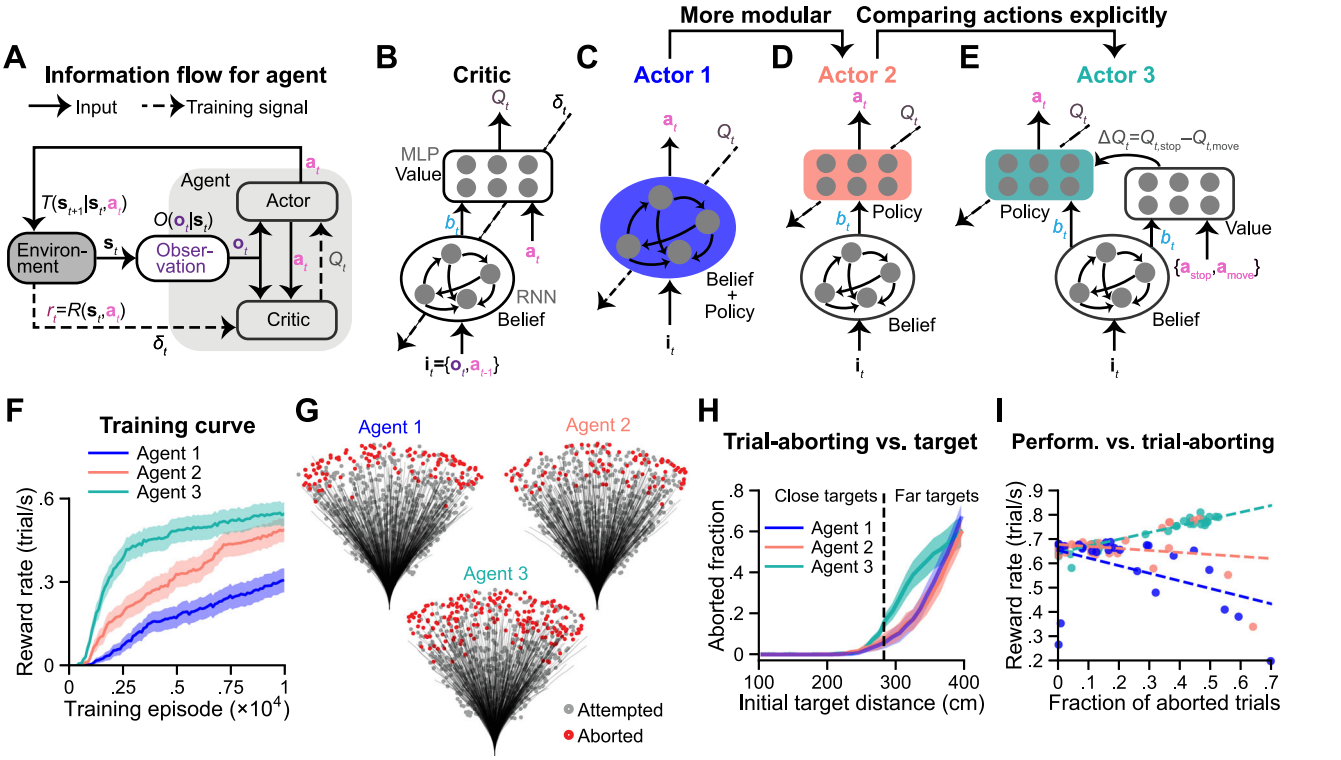


Fig. 2 | Modular RL actors augmented with specific model-free critic values strategically abort trials to maximize reward rate. A Block diagram showing the interaction between an RL agent and the task environment. At each time step t, the agent receives a partial and noisy observation $\mathbf{o}_t$ of the state $\mathbf{s}_t$. The actor network integrates the observation with past observations (via the current belief) and then outputs an action $\mathbf{a}_t$, which partially updates $\mathbf{s}_t$ to $\mathbf{s}_{t+1}$ (the state also has its own dynamics). The actor is trained to output the action that maximizes the value Q_t according to the critic, while the critic is trained to generate more accurate value estimations using the TD error δ_t after receiving reward r_t from the environment (Eq. 2). **B** Schematic of critic. A critic network comprises two modules. The belief module is implemented as an RNN, integrating available information $\mathbf{i}_t = (\mathbf{o}_t, \mathbf{a}_{t-1})$ to compute a belief $\boldsymbol{b}_t$ about the state. The value module outputs the value Q_t of an action $\mathbf{a}_t$ given $\boldsymbol{b}_t$. Both modules are trained end-to-end through back propagation. **C** Schematic of Actor 1 implemented with a single RNN. Given $\mathbf{i}_t$, Actor 1 needs to compute both belief and action together. **D** Schematic of Actor 2, which uses the belief module trained by the critic in (B) to construct b_t. Then, a policy module computes an action $\mathbf{a}_t$ from its input b_t. Only the parameters in the policy module are updated in actor training. **E** Schematic of Actor 3. Similar to (**D**), this network uses the belief module from (**B**) to construct b_t. However, it also directly receives $\Delta Q_t = Q_{t,\text{stop}} - Q_{t,\text{move}}$ from the critic MLP. Actor 3 takes an action $\mathbf{a}_t$ given both b_t and ΔQ_t. **F** Reward rate (rewarded trials per second) during training, measured using a validation set (500 trials) for each agent at each checkpoint (every 100 training episodes). Shaded regions denote ±1 s.e.m. across 50 training runs with different random seeds. **G** Overhead view of the spatial distribution of 800 representative targets and trajectories for Agents 1, 2, and 3 (sampled from random seeds that yielded reward rates >0.2) navigating toward these targets. Red/gray dots denote aborted/attempted targets, as Fig. 1. **H** Fraction of aborted trials versus initial target distance, averaged across random seeds. Shaded regions denote ±1 s.e.m. For each seed, target distance is binned for every 10 cm, and the aborted fraction is calculated in each bin. The black dashed line denotes the median distance ($200\sqrt{2}$ cm) of the targets. **I** Reward rate versus aborted fraction for 2000 targets. Each dot denotes a training run with a unique random seed. Dashed lines denote the linear regression lines for the dots. Source data are provided as a Source Data file.

the actor or inherited from the critic via the output of the critic's RNN. All agents use the same modular critic architecture (Fig. 2B) and are trained with identical objectives for both the critic and actor. Within this common framework, we explore three different actor architectures (Fig. 2C–E).

Actor 1 (Fig. 2C) has a simple, holistic architecture. Given the state-related inputs $\mathbf{i}_t$, it uses a single RNN to compute both its own belief and action, thus lacking architecture-enforced neural specializations. Actor 2 (Fig. 2D) and Actor 3 (Fig. 2E) inherit the critic's RNN belief module. Using the output of this module, b_t, they compute the actions $\mathbf{a}_t$ via an MLP, referred to as the policy module. This design increases modularity by imposing a hierarchical computation—belief first, policy later—rather than the mixed computation used by Actor 1. For Actors 2 and 3, the belief is not computed twice, but simply computed via the RNN (allowing for the integration of velocity observations into position beliefs) in the critic, and then inherited for the actor. We refer to this RNN as the critic's RNN because it is trained through backpropagation on the critic's objective.

Actor 3 is further granted an additional input: the difference in value ΔQ_t (evaluated by the critic's MLP) between stopping ($\mathbf{a}_{\text{stop}}$) vs. executing a movement and thus continuing the trial ($\mathbf{a}_{\text{move}}$):

$$\Delta Q_t = Q_{t,\text{stop}} - Q_{t,\text{move}} \tag{3}$$

In other words, the policy module in Actor 3 (Fig. 2E, green) is not only influenced by b_t from the critic's belief RNN (as is Actor 2), but also by an explicit value signal from the critic's value module. In this sense, the critic's role in actor learning is two-fold. For all actors, it serves as the learning objective to adjust the actor's weights, a slow process[9] characteristic of model-free RL. For Actor 3, it additionally serves as an internal model capable of evaluating actions of interest, such as $\mathbf{a}_{\text{stop}}$ and $\mathbf{a}_{\text{move}}$, to influence the policy unit (MLP) dynamics (a fast process[9], see *Discussion* for more detail): a characteristic of model-based RL.

With the same critic architecture, we refer to RL agents using Actor 1, 2, or 3 as Agent 1, 2, or 3, respectively. Among these agents, Agent 3 exhibited the fastest learning rate and the highest final

performance (Fig. 2F, Fig. S3A). It also had the most training instances (i.e., random seeds), achieving proper learning and showing the development of aborting behavior (Fig. S3B, C, D). Indeed, the behavior of these agents showed that while all networks could abort trials (see Fig. 2G for example sessions), this was more frequent in Agent 3 than Agents 1 or 2 (Fig. S3D, K-S test: Agents 1 vs. 3, $p = 0.018$; Agents 2 vs. 3, $p = 0.002$). Furthermore, the spatial distribution of aborted trials differed: Agent 3 aborted trials for closer distances than the other agents (Fig. 2H, S3E; the exact shape of boundaries dictating when models aborted trials as a function of distance depends on observation noise). Most importantly, aborting was only beneficial for Agent 3: reward rate (considering all trials) increased with trial-aborting behavior in Agent 3 (Fig. 2I: Pearson's $r = 0.95$, $p = 5.45 \times 10^{-21}$; slope = 0.28), but not for Agents 1 or 2. Indeed, aborting trials hindered performance in Agent 1 ($r = -0.44$, $p = 0.023$; slope = -0.31) and did not enhance performance in Agent 2 ($r = -0.2$, $p = 0.24$). This suggests that while the aborting behavior was adaptive and improved long-term return for the former (Actor 3), it was not for the latter two. Interestingly, when we double the training duration of Agent 1 (Fig. S3F–I), the negative correlation between aborting behavior and reward rate became less significant (Fig. S3J: $r = -0.22$, $p = 0.2$; slope = -0.14), demonstrating that while aborting is ultimately beneficial in all agents (i.e., with enough training), it develops much faster in Agent 3 than the rest.

Together, these results suggest that strategically aborting trials to increase reward rates (1) requires a value function that accounts for long-term value beyond just a single trial (which our past work neglected[22]), (2) benefits from a modular actor architecture (Actor 2 and 3 vs. 1; also see Fig. S4 and neural specialization analyses in *Methods*), and (3) can be catalyzed by a model-based setting estimating the value of aborting, or skipping a trial and directly providing this information to the actor network (Actor 3 vs. Actor 2). The results are reminiscent of the difference between a slower synaptic-plasticity based RL algorithm in basal ganglia (i.e., the MLPs computing all state-action values; model-free) vs. a faster acting neural dynamics based RL algorithm in frontal cortex (i.e., the MLP in Actor 3 explicitly provided with task-relevant value differentials; model-based).

Value-informed policy extraction explains reward-maximizing trial-aborting behavior

Next, we aimed to understand the mechanism by which Agent 3 uses its additional learning signal to increase its reward rate by aborting trials. While poor RL performance is often attributed to an imperfect value function[35,36], here Actors 2 and 3 maximize the same type of value functions. In fact, they share the same critic architecture and both inherit beliefs from the critic. The difference between these agents, therefore, ought to be in their actors' abilities to extract an optimized policy from the critic (see[37] for a similar argument). To demonstrate this, we compute ΔQ_t (Eq. 3) for Agents 2 and 3 (see 'Critic's opinion' section in "Methods"). This difference in value is provided as an input to Agent 3, but can also be computed for Agent 2, as any critic can estimate values for the "move" and "non-move" (i.e., stop) actions. Figure 3A shows that while the critics of Agents 2 and 3 have similar "opinions" about the value, $\Delta Q_{t=1}$, of aborting (red) or attempting (gray) trials when target is presented (Agent 2: top, vermilion; Agent 3: bottom, green), the actor's behavior in Agent 3 followed its critic's opinion more closely. This is further illustrated by contingency tables (Fig. 3B) and the fraction of trials in which the actor attempted a trial despite the critic's suggestion to abort (Fig. 3C, Agent 2: 46%, Agent 3: 17%).

To further understand the mechanism by which policies may be extracted for a given "opinion", we may also examine the behavior of populations of artificial units within Agent 3's actor—the network that aborts trials to maximize the reward rate. This actor contains a belief module (Fig. 3D, orange) and a policy module (Fig. 3D, green). At trial onset, $\Delta Q_{t=0}$ is matched (and negative) for subsequently aborted (red, Fig. 3E) and attempted (gray, Fig. 3E) trials. In other words, when the target is first presented, the value of moving is higher than the value of stopping ($\Delta Q_{t=0}$ is negative). This is because to start (and then potentially abort) a trial, the agents first have to move. If they do not move, the trial will last until a time-out, akin to the monkey's "never start" trials (see 'Reward' section in "Methods"). Immediately after the initial step, however, $\Delta Q_{t=1}$ becomes positive ($Q_{t,\text{stop}} > Q_{t,\text{move}}$) for subsequently aborted trials but not for attempted ones. Thus, we may reason that if units were reflecting ΔQ_t, there should be a divergence in neural activity for subsequently aborted vs. attempted trials already at the second time-point after targets (i.e., offers) are presented.

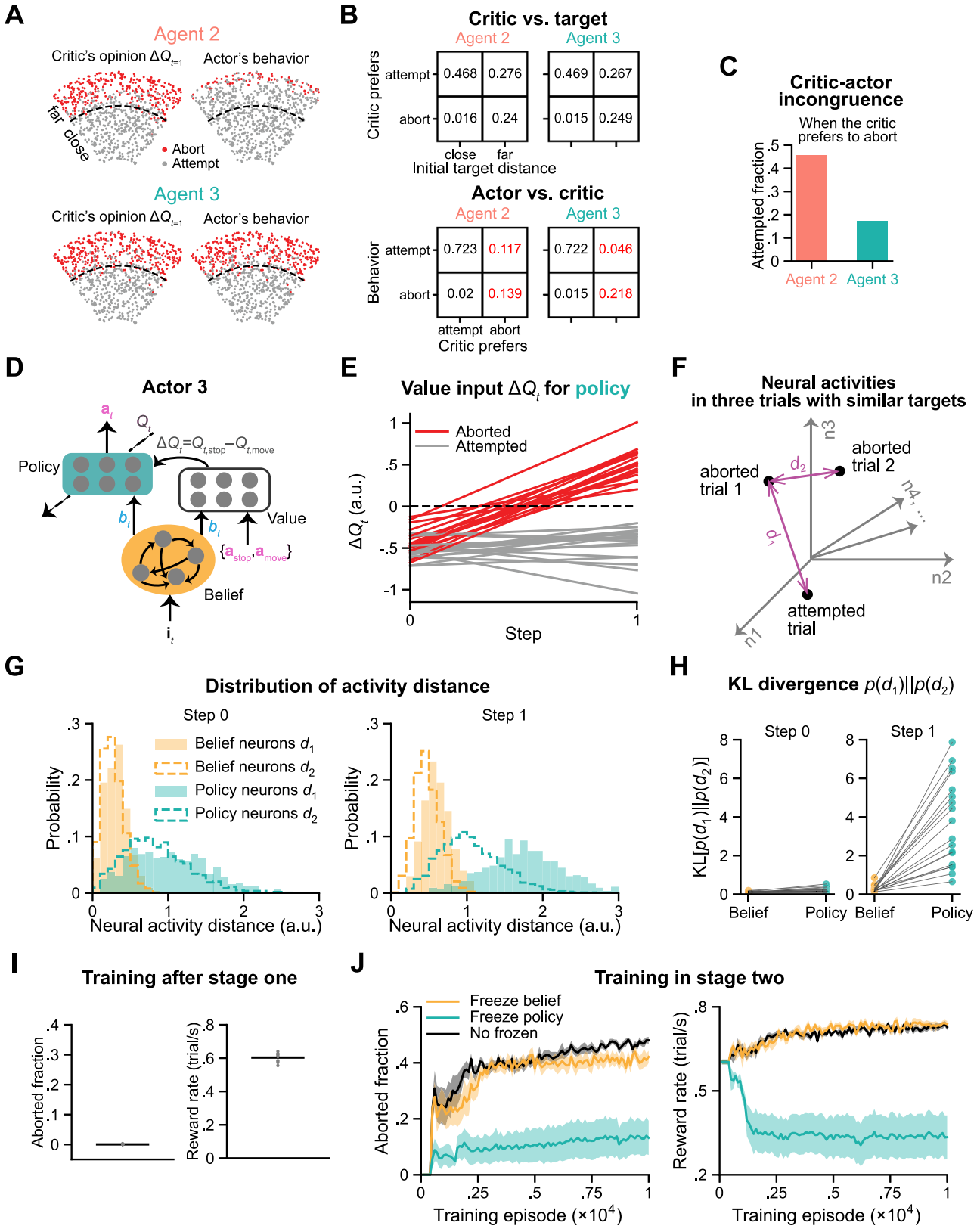
To examine this hypothesis, in Agent 3 we matched each aborted trial with (1) its most similar attempted trial (in terms of targets' Euclidean distance upon presentation) and (2) its most similar aborted trial. Then, we expressed the artificial neural network's population activity in a high-dimensional space (Fig. 3F) and compared the Euclidian distance between aborted and attempted trials (d_1), as well as the distance between two aborted trials (d_2) at trial onset/target presentation ($t = 0$) and the immediately subsequent timepoint ($t = 1$) in the belief (orange) and policy (green) modules. We repeated this procedure for all aborted trials in a 2000 trial session to estimate full distributions of these distances (Fig. 3G). Results showed that, at trial onset, belief and policy neurons do not differentiate between attempted and aborted trials more than they do between pairs of aborted trials (Fig. 3G, left). In contrast, after the initial step, the neural state difference between attempted and aborted trials (d_1) is distinct from that between two aborted trials (d_2) within the policy module, but not in the belief module (Fig. 3G, right). This effect can be quantitatively assessed by calculating the KL divergence between the distributions of d_1 and d_2 (Fig. 3H).

To further verify the role of the belief and policy modules, we conducted ablation analyses by freezing neurons during the learning of trial-aborting behavior (see 'Agent Training' section in "Methods"). We first trained agents to navigate to targets without considering long-term outcomes, i.e., only considering the target in the current trial (training stage one; as in our prior work[22]). Namely, agents attempted to maximize

$$\mathbb{E}_{\mathbf{s}',\mathbf{a}'}\left[\sum_{k=0}^{N-t}\gamma^k r_{t+k}|\boldsymbol{s}_t,\boldsymbol{a}_t\right] \tag{4}$$

where N denotes the final step in a trial. In this stage, agents did not abort trials (Fig. 3I). Then, in stage two, we reverted the definition of value to that in Eq. 1 by changing the upper bound of the sum from $N-t$ to ∞: we now considering not only the current target up to time N, but also all times for all subsequent targets. We continued training the agent while either (1) freezing parameters in the belief module (Fig. 3D, orange), (2) freezing parameters in the policy module (Fig. 3D, green), or (3) allowing all parameters to be updated. The results show that preventing the belief module from learning during the second stage of training impacted neither the agent's trial-aborting behaviors (Fig. 3J, left, orange) nor performance (Fig. 3J, right, orange) compared to the agent with no frozen parameters (Fig. 3J, black). In contrast, freezing the policy module (Fig. 3J, green) during stage two training resulted in very poor behavior.

Together, these results suggest that reward-enhancing aborting behavior is supported by a modular actor network whose policy module is conditioned on a specific value learning signal allowing for better (or faster) alignment with the critic: optimization of policy extraction. Further, this reward enhancing behavior is supported by the policy module but not the belief module of the actor, as evidenced by artificial neural activity at target presentation, and confirmed through ablation (neuron-freezing) analyses.



Selective units in dorsolateral prefrontal cortex distinguish between aborted and attempted trials

The RL modeling suggests that session-long aborting behaviors may be learned faster by a policy module directly informed of value differentials between performing the task at hand or not. This difference is evident in the agent's neural activity already at the time-point immediately following target presentation. The modeling work also suggests a functional specialization between units coding for key task-relevant *beliefs* (e.g., distance-to-target) and the *policy* to be employed given that belief (i.e., belief neurons do not reflect upcoming aborts). We next investigated whether these patterns occurs in real neural activity of macaques performing the "firefly task"[20,21,24–26].

We recorded single-unit spiking activity from the dorsolateral prefrontal cortex (dlPFC; 445 single units), parietal area 7a (2647 single

Fig. 3 | Policy neurons reflect a trial-aborting strategy while belief neurons do not. A Overhead view of the spatial distribution of 800 representative targets for a training run of Agent 2 (top) or Agent 3 (bottom). Red/gray dots on the left denote that agents' critics valued abort/attempt more ($\Delta Q_{t=1} > 0$ / $\Delta Q_{t=1} < 0$), while on the right they denote whether trials were aborted/attempted based on behavior. Black dashed arcs equally separate the target sampling area by target distance. **B** Numbers in the tables denote fractions of total trials. *Top*: Contingency tables showing the critic's preference ($\Delta Q_{t=1}$) vs. initial target distance. *Bottom*: Table showing the actor's behavior vs. the critic's preference. Red color highlights pronounced differences between agents 2 and 3. Trials from all random seeds (with reward rate > 0.2) were included. **C** Fraction of attempted trials (the red number in the first row of each table in **B**) among trials that the critic preferred to abort (column sum of red numbers in each table in **B**) for two agents. **D** Schematic of Actor 3, shown as in Fig. 2E. We use yellow to denote beliefs, both for the module in this panel and for traces in subsequent panels. **E** ΔQ_t computed by the value module in **D** during the first two timesteps of trials. 2000 trials were used to evaluate the agent with each random seed. Each line denotes ΔQ_t either averaged across aborted trials (red) or across target-matched attempted trials (gray) for each seed. We use a.u. to denote arbitrary units. **F** Illustration of measuring distances in high-dimensional neural activity space between activities for an aborted and a target-matched attempted trial (d_1), or between activities for two target-matched aborted trials (d_2). **G.** Distributions of d_1 and d_2 at timestep 0 (left) and timestep 1 (right) for belief and policy neurons in an agent trained with a random seed. **H** KL divergence between distributions of d_1 and d_2 in (**G**), including all random seeds (each dot denotes a random seed). **I** Aborted fraction and reward rate after training stage one, averaged across training instances. 500 trials were used. Gray dots: 10 training instances. **J.** Fraction of aborted trials (left) and reward rate (right) over stage two training. After training stage one (I), the agent then continued to train with either belief or policy neurons frozen, or with none of the neurons frozen. 500 trials were evaluated for each agent at each checkpoint, occurring every 100 training episodes. Shaded regions denote ±1 s.e.m. across training instances. Source data are provided as a Source Data file.

units), and the dorso-medial superior temporal area (MSTd; 240 single units). In prior work we have solely analyzed "attempted trials" (Fig. 1B, leftmost), and established that coding of the key latent variable within this task (i.e., the instantaneous and unobservable distance to target) is strikingly distributed (e.g., is decodable from dlPFC, area 7a, and MSTd[25]) and involves recurrence, including bi-directional coupling between sensory and cognitive areas[24]. This distributed and recurrent nature of the coding of the key latent variable (i.e., "belief") is conceptually akin to the RNN computing beliefs in our RL agents. Here, instead, we aim to establish whether we can identify biological units matching the properties of the agent's policy module (e.g., MLP in Actors 2 and 3, Figs. 2 and 3).

As for the analysis of the artificial units (Fig. 3F–H), we matched each aborted trial to its most similar attempted trial in terms of target distance, eccentricity, and afforded reward rate. Further, we only examined neural responses during target presentation (−0.3 to 1 second post-stimulus onset): within this window, motor output is on average not yet different between attempted and aborted trials. Indeed, Fig. 4A shows the subsampled distribution of matched target locations for attempted (gray) and aborted trials (red; top row) for three example sessions (one per monkey), as well as velocity profiles for example trial matches (second row) within the time period of interest (see Fig. S5 for additional example sessions and trials). Next, we constructed raster plots (Fig. 4A, third and fourth row) and peri-stimulus time histograms (PSTHs; Fig. 4A, fifth row) for neurons in each region, based on the subsampled set of trials we could match. We found neurons in MSTd (Fig. 4A, leftmost), area 7a (Fig. 4A, middle), and dlPFC (Fig. 4A, rightmost) that seemingly differentiated between attempted and aborted trials (Wilcoxon sign-rank test, baseline-corrected spike counts between 0 and 0.3 sec after stimulus onset; $p < 0.05$; figures show example neurons, one per monkey). Responses from these neurons were heterogenous, but importantly, they did not merely show a lack of response during aborted trials, as if the animals had 'missed' the target presentation. On the contrary, on occasion (e.g., Fig. 4A, dlPFC example, Monkey M) the response to aborted trials was greater than that to attempted ones. Further, the vast majority of neurons that differentiated between subsequently aborted or attempted trials did so within the first 300 ms following target (i.e., offer) presentation.

A greater fraction of neurons in dlPFC differentiated between policies (i.e., attempt vs. abort) for matched offers compared to the other recorded areas. Specifically, 12.6% of dlPFC neurons (Fig. 4B, red; computed as the average fraction within each session and then taking the median across sessions) distinguished between attempted and aborted trials based on baseline-corrected spike counts measured 0–0.3 s following stimulus onset ($p < 0.05$). This fraction was significantly greater than that observed in area 7a (4.3%; Mann–Whitney U, $p = 7.8 \times 10^{-5}$) and MSTd (0.5% median; $p = 7.7 \times 10^{-7}$), and was also significantly above chance ($p = 0.02$; chance level = 5%). In contrast, neither area 7a ($p = 0.53$) nor MSTd ($p = 1.0$) exhibited a greater-than-chance fraction of neurons sensitive to policy during target presentation (Mann–Whitney U test). Although the median fraction of MSTd neurons differentiating aborted from attempted trials was near zero across sessions, this primarily reflects the fact that in most sessions no MSTd units exhibited such selectivity. Importantly, in a subset of sessions ($n = 12$), some MSTd neurons did differentiate between aborted and attempted trials, with selectivity emerging within the first second following stimulus presentation (see Discussion). These results across MSTd, area 7a, and dlPFC remained unchanged when analyses were performed without first averaging within sessions (which may be sensitive to sample size), instead pooling neurons across all recordings and directly computing medians (dlPFC = 9.61%; 7a = 5.48%; MSTd = 3.42%; $\chi^2 = 13.79$, $p = 0.001$). We further restricted this analysis to sessions with acute recordings only. Indeed, MSTd, as a deep structure, was sampled exclusively using acute silicon probes, whereas dlPFC and area 7a were sampled using both acute silicon probes and chronic Utah arrays. Even under this restriction, dlPFC continued to exhibit a greater fraction of neurons differentiating aborted from attempted trials (dlPFC = 9.21%; 7a = 4.24%; MSTd = 3.42%; $\chi^2 = 12.23$, $p = 0.00221$). Finally, these findings were robust to additional analysis controls. Restricting analyses to trials in which animals were immobile prior to target presentation (Fig. 4C; linear speed = 0 cm/s, 88% of trials) yielded the same result (dlPFC fraction vs. chance, $p = 0.018$). Likewise, when further restricting analysis to trials in which aborting did not occur until at least 1 s after stimulus offset (Fig. 4D; 67.5% of trials), dlPFC neurons still differentiated between policies at a rate exceeding chance ($p = 0.007$).

Lastly, we asked whether neurons that differentiated between attempted and aborted trials preferentially encoded specific task variables, and whether they were more specialized or multiplexed in their encoding of task variables. To do so, we first fit a fully coupled Poisson generalized additive model (pGAM) to each neuron to estimate tuning to sensorimotor, visual, and latent "belief" variables. As expected from prior work[24,25], area 7a showed the strongest encoding of sensorimotor variables such as linear and angular velocity and acceleration, MSTd was most strongly driven by eye-movement dynamics (likely reflecting associated optic-flow input), and dlPFC showed the highest prevalence of tuning to latent task variables, including the animal's continuous and unobservable distance to the target (Fig. 4E; $\chi^2 = 112.7$, $p < 10^{-24}$ for differences in tuning prevalence across areas). Importantly, all three areas exhibited tuning to latent belief variables, consistent with the distributed and recurrent representation expected from an RNN-like computation (Figs. 2, 3). On average,

https://doi.org/10.1038/s41467-026-74783-6

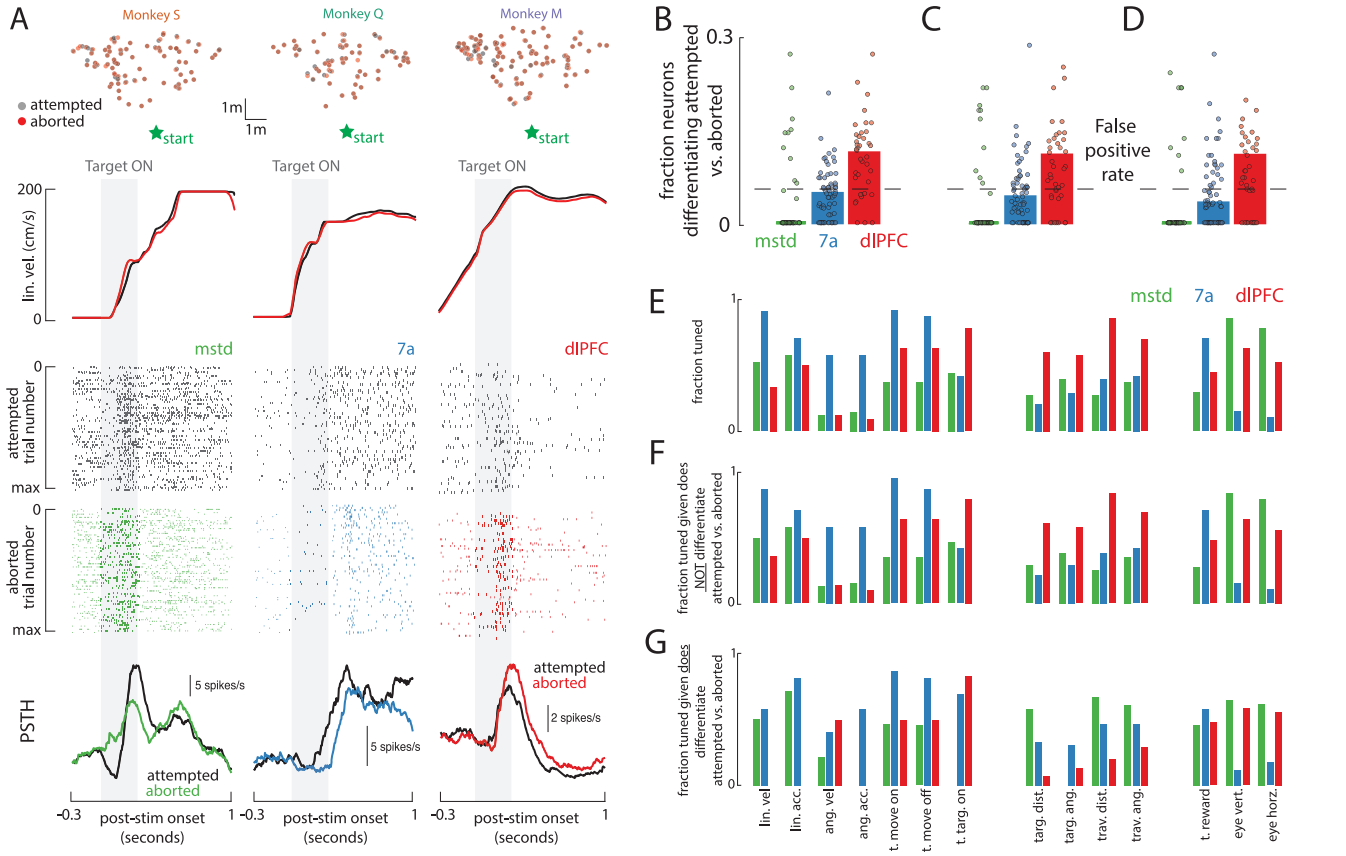


Fig. 4 | Dorsolateral prefrontal cortex reflects differing policies. A Data from 3 example cells, one from each animal and brain area: MSTd (green; 1st column), area 7a (blue; 2nd column), and dlPFC (red; 3rd). Top row shows the distribution of matched target locations (overhead view) for attempted (gray) and aborted (red) trials. Starting location of the animal is shown by a green star. The second row shows the linear velocity profile for an example matched attempted (gray) and aborted (red) trial. This demonstrates that in the period surrounding target presentation (−0.3 to 1 second post-stimulus onset), the velocity profiles can be matched across trial types. The transparent gray area shows the period of time the target was visible (0 to 300 ms after stimulus onset). Third and fourth row respectively show example raster plots for attempted (black) and aborted trials (colored as a function of brain area). Each row is a trial and each dot is a spike. Fifth row shows the peri-stimulus time histogram (PSTH) of the data presented in the raster plots (aborted in black and attempted as a function of brain area). **B** Fraction of neurons differentiating in their peri-stimulus time response (spike count from 0 to 300 ms after stimulus onset, baseline corrected) between attempted and aborted trials. Each dot is the fraction computed on individuals sessions. Green (MSTd), blue (7a), and red (dlPFC) bars are the median across sessions. Data from MSTd (due to its lower total yield) is not Gaussian so we performed non-parametric significance testing (Mann-Whitney U test). Dashed black line shows the false positive rate at 5%, given the significance threshold employed. **C** Same as (**B**), when only considering trials where the animal was stopped (linear velocity = 0 cm/s and angular velocity = 0 deg/s) before stimulus onset (−300 to 0 ms). **D** Same as **C** when additionally filtering out trials wherein the animal aborted within 1 sec of stimulus onset. **E** Fraction on neurons tuned to task variables in MSTd (green), area 7a (blue), and dlPFC (red) as estimated by the pGAM. We included a total of 14 task variables. Sensorimotor variables (first cluster) included linear velocity, linear acceleration, angular velocity, angular acceleration (continuous variables), as well as the time of movement onset, offset, and target presentation (discrete variables). Latent ("belief") variables (second cluster) included the current radial distance to target, angular distance to target, distance traveled, and angular distance traveled (all continuous variables). Finally, "other" variables (third cluster) included the time of reward presentation (discrete) and the vertical and horizontal eye positions (continuous). **F** as **E** restricted to neurons that did not differentiate between attempted and aborted trials in the 300 ms following target presentation. **G** as **F** restricted to neurons that did differentiate between attempted and aborted trials. Source data are provided as a Source Data file.

each neuron was tuned to 5.6 (± 0.15) of the 14 task variables included in the pGAM.

We next asked how these tuning distributions changed when neurons were split according to whether they differentiate attempted from aborted trials (Figs. 4F and 4G). As expected, given that most neurons across all areas did not significantly differentiate policies, the tuning distributions of non-policy-selective neurons closely resembled the overall pattern (Kruskal–Wallis test across areas for all covariates: $H = 1.82$, $p = 0.40$; Fig. 4E vs. Figure 4F). In contrast, the subset of neurons that did differentiate between attempted and aborted trials (5.1% of neurons overall, most in dlPFC) showed a distinct pattern of tuning (Fig. 4G). Notably, while area 7a and MSTd respectively remained the areas most frequently tuned to sensorimotor and eye-related variables, dlPFC was not the area most commonly tuned to latent variables among policy-selective neurons ($\chi^2 = 6.94$, $p = 0.031$, comparing latent-variable tuning fractions across MSTd and dlPFC). Further, while neurons differentiating between attempted and aborted trials in area 7a (attempted = 6.3 ± 0.26; aborted = 6.2 ± 0.24) and MSTd (attempted = 4.9 ± 0.71; aborted = 5.1 ± 0.97) were tuned to the same total number of task variables (both $p > 0.46$, Mann–Whitney U), dlPFC neurons differentiating between attempted and aborted trials were tuned to significantly fewer task variables (3.1 ± 0.74) than dlPFC neurons not differentiating between attempted and aborted trials (5.6 ± 0.72; $p = 0.015$). That is, dlPFC neurons seemingly indexing policy were less likely to code for other task variables.

Together, these results suggest that (1) dlPFC contains the largest fraction of neurons differentiating between attempted and aborted trials, (2) this differentiation emerges already during offer presentation—well before the animal expresses aborting behavior, as suggested

by the reinforcement learning modeling, and (3) dlPFC neurons signaling policy are significantly more specialized in their tuning, particularly by *not* encoding latent beliefs, compared to neurons in the same or other areas that do not signal policy.

Dorsolateral prefrontal cortex population activity precedes aborting behavior and correlates with adaptive policy optimization

To examine population-level responses (as opposed to a series of individual neurons, Fig. 4) we first performed the same analysis as for the model in Fig. 3F–H. That is, simultaneously recorded neural activity from a given area defined a high-dimensional space (average firing rate 0 to 300 ms post-stimulus onset), with each neuron representing an independent axis. Then, we paired each aborted trial with two other trials from the same session that had the most similar targets, one attempted and another aborted trial. We computed the Euclidian distances within these high-dimensional spaces, and contrast the distance in neural activity between pairs of aborted-attempted trials (d_1) or pairs of aborted trials (d_2) for each brain region (akin to Fig. 3F). As shown in Fig. 5A, the results demonstrate a separation in neural space between attempted and aborted trials that exceeds that for paired aborted trials in dlPFC (red), but not 7a (blue) or MSTd (green). This effect can be quantified by calculating the KL divergence between distributions of d_1 and d_2 for each brain region, which demonstrated greater divergence in dlPFC than MSTd or area 7a (Fig. 5B, one-way ANOVA, $p < 0.001$; MSTd vs. 7a, $p = 0.68$). Further, examining neural activity along the top principal component within either dlPFC, 7a, or MSTd, we corroborated that only dlPFC distinguishes between attempted and aborted trials upon target presentation (Fig. 5C). This separation occurs in dlPFC on average 291 ms post-stimulus onset, a timing that comfortably precedes the average duration from stimulus presentation to abort (1661 ms; comparison of timing of neural divergence vs. abort across sessions, $p < 0.001$).

Next, to provide a quantitative metric of the variance explained by the neural activity encoding aborting behavior, we performed a demixed principal component analysis (dPCA[38]). This method is conceptually akin to PCA, but keeps the learned components (or "subspaces") interpretable vis-à-vis task features, such as aborts. Figure 5D shows neural activity of an exemplar session in each area, projected first onto a behavior-naïve subspace (Fig. 5D, top), and then onto a subspace differentiating between aborted (colored by brain area) and attempted trials (Fig. 5D, bottom). Mimicking the PCA analyses, we observe that dlPFC but not the other areas differentiates between aborted and attempted trials, and that this bifurcation occurs shortly after stimulus presentation (278 ms) and well before the monkey aborts ($p < 0.001$). On average, the subspace accounting for aborting behavior accounts for 21.9% of the variance accounted by the behavior independent subspace in dlPFC, and 10.2% and 11.3% respectively in MSTd and area 7a (Fig. 5E, one-way ANOVA, $p = 0.002$).

Lastly, to examine whether and how this population activity relates to adaptive behavior, we quantified how aborting impacts reward rates for each animal and session, and then correlated this session-wise measure of abort adaptiveness with the variance explained by the neural subspace encoding subsequent aborts (Fig. 5E). To compute adaptiveness, behavioral simulations compared three policies: (1) the animals' actual behavior, (2) a counterfactual policy in which animals never aborted, and (3) a theoretical upper bound on achievable reward rate (see "Methods"). This analysis showed that aborting was indeed adaptive overall (actual vs. no-abort policy, 0.154 rewards/s vs. 0.11 rewards/s; $p < 0.001$), and that this held for each animal individually (Fig. 5F, colored, all $p < 0.01$). Nonetheless, all animals remained below their theoretical maximum reward rate, indicating substantial room for further optimization (actual vs. optimal policy, 0.154 rewards/s vs. 0.214 rewards/s; $p < 0.001$). Interestingly, the degree to which aborting increased reward rates on individual sessions (i.e., actual−no aborts) correlated with the variance explained by aborting subspaces in dlPFC ($r = 0.45$, $p = 0.0046$; Fig. 5G). This was not true for MSTd ($r = -0.19$, $p = 0.27$) or area 7a ($r = -0.02$, $p = 0.86$).

Discussion

We developed a task wherein animals naturalistically forage for reward in virtual reality[19–22,24–26,39]. A central feature of this task, just as in real-world foraging, is the implicit ability and importance of *manipulating* (i.e., minimizing) the time between possible rewards. This breaks with the mold of most traditional systems neuroscience work which, by and large, experimentally determines the speed at which an agent can acquire rewards (e.g., the speed of stimulus presentation and the duration of inter-trial intervals). Similarly, in traditional approaches, animals usually have a very limited ability to impact the structure of their tasks: for example, how valuable the probable next offer is relative to the current (see[40–46] for recent exceptions using 'time' as a key experimental variable to estimate value[40,41] or confidence[42–46]). Instead, in our task, animals can abort trials, saving themselves valuable seconds. This behavior, however, requires completely forgoing the potential for reward on the current trial.

Using this task, here we demonstrate (1) that macaques have knowledge about their own performance and strategically use this knowledge in aborting trials in a session- and individual-specific manner to increase (but not fully maximizing) session-long reward rates, (2) that computationally this behavior requires considering state-action values beyond a single trial, and (3) that this behavior is learned faster with a modular actor-critic architecture in which a policy module is given an explicit estimate of the relative value of aborting the current offer. This explicit signal (4) makes it easier for the actor to use the critic's evaluations (i.e., policy extraction) within the actor-critic architecture. Moreover, we show that (5) individual neurons and (6) population dynamics in dlPFC but not parietal area 7a or MSTd reflect the strategic aborting of trials upon target presentation—even if the aborting itself won't occur for another second or longer. Lastly, (7) the fact that a putative policy module (i.e., "dlPFC") reflects subsequent aborting behavior already during target presentation is as predicted by the modular actor-critic artificial neural network, as is (8) the fact that belief and policy modules/neurons are relatively specialized: while they index subsequent aborting, they overall index fewer task variables.

Previously, we have shown that coding of distance to target—the key latent variable that must be minimized in this task—is distrusted and recurrent[24,25]. That is, single units across cognitive and sensory cortices (i.e., dlPFC, 7a, and MSTd) are tuned to the instantaneous angular and/or radial distance to target[24]. Dorsolateral prefrontal cortex and MSTd appear to have a preference for distal targets, and preferentially couple with one another when animals keep track of the (invisible) target position with their gaze[24] (i.e., an embodied strategy). Parietal area 7a is more frequently active as targets get closer[24] and is the area most clearly encoding sensorimotor state variables (e.g., linear velocity). Likewise, distance to target is decodable from all three of these areas[25]: it is more invariant to sensory context in dlPFC and 7a, and is more modulated by sensory observations in MSTd. Together, these previous results suggest that MSTd and 7a are strongly modulated by observations and state variables (also see[26,47]), while the encoding of beliefs are more distributed. Interestingly, in the current study we see clear area specialization within this "firefly task": dlPFC but not other areas reflect the animal's policy. We also show that these "policy neurons" in dlPFC are distinct from putative "belief neurons" (Fig. 4E-G).

More broadly, these results add to our understanding of reinforcement learning within frontal, parietal, and sensory circuits, and beyond the striatum. Much of the existing neuroscience literature has focused on tasks in which value can be approximated as a state value,

https://doi.org/10.1038/s41467-026-74783-6

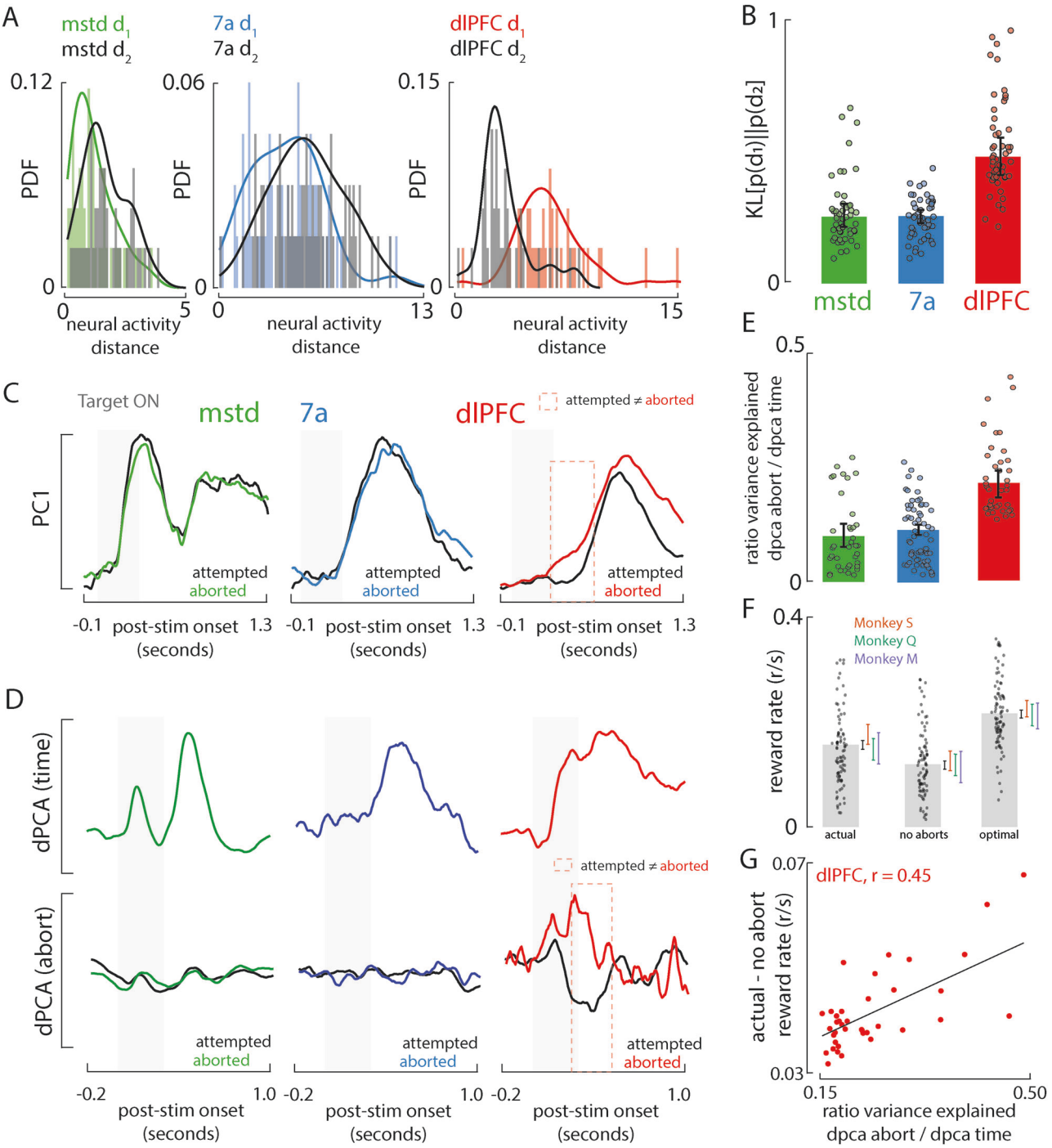


Fig. 5 | Population dynamics and their relation to behavior. A Probability density functions of d_1 (colored; Euclidean distance in high-dimensional space between matched attempted and aborted trials) and d_2 (black; distance between matched aborted trials) in MSTd (green), area 7a (blue), and dlPFC (red). **B** KL divergence between distributions shown in (**A**). For statistical testing, we subsampled (without replacement) 50 times (dots) from each distribution, for 20 sessions. Bars shows the means across subsamples and error bars are S.E.M. **C** Principal component of neural activity in each area, separated by attempted (black) and aborted (colored) trials. Red dashed box shows time-period over each the first PC in dlPFC differentiates between aborted and attempted trials. **D** De-mixed PCA, showing a condition-independent subspace (time, top), and the subspace differentiating between attempted (black) and aborted (colored) trials (bottom). **E** Fraction of variance explained by the "abort subspace", normalized to the condition-independent subspace. This abort subspace accounts for more variance in dlPFC (red) than the other areas. Bars represent means and error bars are S.E.M. across sessions. **F** Reward rate (rewards/s) that we actually observed vs. counterfactuals of what the reward rate would have been if no aborts were made (no aborts), or if animals optimally aborted (optimal). Each dot is a session, black error bar is mean across animals, and colored error bars correspond to individual animals means and their S.E.M. (S, Q, or M). **G** Correlation between the increased in reward rate afforded by aborting trials (y) and the variance explained by the abort subspace in dlPFC across session. Source data are provided as a Source Data file.

with the ventral striatum playing a prominent role in outcome prediction and reward prediction errors[48–50]. In more naturalistic and continuous tasks, however, value depends on both state and action, and on their consequences over extended time horizons. Prior work suggests that the dorsal striatum is particularly involved in action selection and the learning of state–action contingencies, integrating cortical representations of candidate actions with value-related signals[51–53]. Rather than computing explicit state–action values in isolation, dorsal striatal circuits are thought to bias policy execution by modulating the likelihood of initiating or sustaining particular actions. In this framework, dlPFC provides structured representations of task goals, action plans, and long-range policies, which are iteratively shaped through interactions with striatal value-sensitive gating mechanisms. Interpreting our favored model (Agent 3) at this functional level, the value module of the critic captures a role analogous to dorsal striatal evaluation of action desirability, while the policy module reflects dlPFC involvement in policy selection and arbitration (Fig. S7A). Neurobiologically, such interactions are implemented through closed cortico–basal ganglia–thalamo–cortical loops, as the dorsal striatum influences dlPFC indirectly via the pallidum and mediodorsal thalamus (Fig. S7B). These relay nodes are therefore likely to participate in conveying value-dependent signals that shape policy selection, rather than transmitting explicit value estimates alone.

We also show that allowing the value module to project values for "stopping" and "moving" actions to the policy module improves learning rates and task performance by allowing animals to abort low reward rate offers. Importantly, this added signal to the policy module does not change value computations, but allows for an optimized policy extraction (i.e., critic and actor becoming more congruent with one another). In this same vein, prior studies have shown that striatal medium spiny neurons expressing D1 or D2 dopamine receptors, which respectively facilitate or inhibit actions, correlate positively and negatively with state-action values[54,55]. These results support the argument that D1 and D2 striatal neurons first decide whether or not to initiate an action, with specific actions being selected thereafter[55,56]: perhaps not so dissimilar from "coarse-to-fine" mechanisms known in sensory cortices, with a first step deciding whether to engage in the trial at all or not, before picking finer actions. Our results align with this view, suggesting that the dlPFC may modulate the balance between D1 and D2 neurons in deciding whether or not to engage with a particular offer. Seemingly, this two-step approach (first decide whether to act, then decide on the particular action) optimizes policy extraction. This interpretation is also in line with hierarchical RL and the suggestion that dlPFC may play a particularly important role in the arbitration between different and potentially concurrent RL algorithms that may exist in biological networks[56–59].

It is interesting to speculate how the dlPFC signals reported here, which we interpret as policy-related, relate to other well-described dlPFC signals such as decision confidence[11,12,60–63] and "No-Go" activity during action withholding[64–68]. During aborted versus attempted trials, confidence in the current policy, or even in the immediate next step of an action sequence, could modulate whether animals continue or terminate behavior, effectively linking confidence estimation to policy evaluation. While confidence was not explicitly measured here, future versions of this task could incorporate explicit wagering or opt-out reports to dissociate confidence in a trial as a whole (exogenous, explicit post-trial report) from confidence in ongoing action execution (endogenously "reported" in a continuous manner by aborting behavior during the execution of a trial). Decisions to abort also bear a resemblance to classical No-Go paradigms, in which prefrontal cortex encodes instructed response suppression[64–68]. However, we speculate that the signal reported here is not an endogenous analog of the exogenously driven No-Go pre-frontal activity, given the substantial delay between dlPFC differentiation of aborted versus attempted trials (~ 300 ms post-stimulus) and the eventual abort (~ 1200 ms). Instead, we speculate dlPFC activity may set a value-dependent baseline or threshold for aborting, with downstream circuits such as the basal ganglia (in particular the dorsal striatum) determining the precise timing of abort.

To further test the putative role of dlPFC as a value-informed policy module, future work should examine dlPFC dynamics during learning of the firefly task, or other tasks that (1) engage model-based reinforcement learning[21,22], (2) allow for multiple policies given matched offers, and (3) index naturalistic closed-loop behaviors in which time-to-reward is manipulable. In addition, systematically varying state-transition statistics (e.g., correlating target distances across trials), while simultaneously recording from candidate cortical and striatal actor and critic modules, would help clarify how value and policy signals are distributed across fronto-striatal loops in continuous action-perception loops. Complementary causal approaches, such as focal lesions or electrical stimulation, will also be essential to establish the necessity and directionality of these interactions. Moreover, given that a central advantage of meta-RL is rapid adaptation to novel environments, future work should investigate how dlPFC supports generalization across tasks and contexts, and how policy representations are reused or reconfigured when task structure changes[14,25,69–71]. Testing multiple variants of the firefly task, or related paradigms, by manipulating sensory and action uncertainty, motor cost, and reward probabilities and magnitudes, among others, would further constrain and orthogonalize model parameter estimation within an inverse rational control (IRC) framework[72], allowing for the fitting of full behavioral trajectories rather than summary statistics alone. This would enable inference of animals' instantaneous beliefs and internal states as actions unfold in real time. Formal model fitting within a framework such as IRC would also directly address an important limitation of the current work. Indeed, here the comparisons between RL models and animal behavior are intentionally qualitative. While we show that a modular actor–critic architecture augmented with an explicit signal encoding the value of engaging in an action (Agent 3) best captures key summary statistics of macaque behavior, substantial discrepancies remain. We emphasize that the computational framework presented here is not intended to achieve fine-grain alignment between RL agents and monkey behavior. The inductive biases of artificial agents and biological systems differ at multiple levels, including the instantiation of neurons (continuous units vs. spikes), architectures (RNNs and MLPs vs. biological circuits), learning rules (backpropagation vs. local synaptic plasticity), and inputs (abstract task variables vs. rich multisensory information). Moreover, artificial agents are optimized for a single task, whereas the brain is shaped by evolution to support many behaviors and is influenced by subjective factors such as fatigue and mood that are absent in standard RL formulations. Accordingly, our goal is neither to build a digital twin of the monkey nor to claim exact behavioral equivalence between agents and animals. Rather, we aim to illustrate a general computational principle: within a model-free RL framework, leveraging a model-based value representation to deliberate between acting and aborting based on the current offer can substantially improve performance. This principle is not specific to any particular system—biological or artificial—but instead reflects a general strategy that, when implemented at a high level, gives rise to meaningful yet qualitative similarities across biological and artificial network behavior and computation. From a modeling perspective, our hybrid architecture illustrates how brain-inspired mechanisms can inform the design of more effective artificial agents. Future work could leverage the IRC framework[72] to account for animal behavior that is suboptimal with respect to the experimenter-defined task objective yet optimal under the animal's subjective goals, potentially enabling more quantitative alignment between artificial agents and animal behavior.

In conclusion, we demonstrate that macaques may reason about their own long-term performance and idiosyncratically decide what

and when experimental offers are worth working for, and which are not. Their aborting of trials is not random, but instead acts to increase reward rates (i.e., they abort low reward rate offers). This computation is best recapitulated by a modular actor-critic architecture in which a policy module is explicitly informed of the relative value of performing the task, or not. This additional signal improves the use of value estimates (i.e., policy extraction), and leads the policy module to distinguish between subsequently aborted and attempted trials already during offer (i.e., target) presentation. We observe a qualitatively similar behavior for individual neurons and populations in dlPFC, but not in 7a or MSTd in macaques.

Methods

Animals, animal shaping, and dataset

Three adult male rhesus macaques (*Macaca Mulatta*; 9.8-13.1 kg) were studied. We analyzed behavioral and neural data from 30 sessions in monkey S, 28 sessions in monkey Q, and 14 sessions in monkey M, resulting in a total of 72 sessions. The animals performed an average of 1355 trials per session, resulting in a total dataset of 97591 trials (range = 449 – 2250 trials). These trials/sessions have been reported on before[24,25], yet in those cases we focused exclusively on attempted trials and disregarded failure modes: aborted trials, "never started" trials, and "never stopped" trials. Here, our focus is in understanding the neural underpinning of aborted vs. attempted trials, which constituted an independent dataset from those previously reported.

Animals were chronically implanted with a lightweight polyacetal ring for head restraint. They were then trained via standard operant conditioning to path integrate to the location of briefly visible targets ("fireflies") by accumulating optic flow velocity signals[42]. All surgeries and procedures were approved by the Institutional Animal Care and Use Committee at Baylor College of Medicine and New York University, and were in accordance with National Institutes of Health guidelines.

Experimental Setup

Monkeys were head-fixed and secured to a primate chair. To navigate, the animals used an analog joystick (M20U9T-N82, CTI electronics) with two degrees of freedom controlling their linear and angular speed in a virtual environment (linear max = 200 cm/s; angular max = 90 deg/s). A DLP projector (3-chip Christie Digital Mirage 2000, Cypress, CA, USA) rear-projected images onto a 60 x 60 cm screen that was attached to the front of the field coil frame, 32.5 cm from the monkey. The projector rendered stereoscopic images generated and rendered using C++ Open Graphics Library (OpenGL; Nvidia Quadro FX 3000 G) by continuously repositioning a virtual camera (60 Hz; height of 1 m) based on the monkey's joystick inputs. The virtual environment was composed of ground plane textural elements with a lifetime of 250 ms (at exception for their first presentation, in which each element had a different lifetime drawn from a uniform distribution between 0 and 250 ms). These elements were isosceles triangles (base x height: 8.5 ×18.5 cm^2) that were randomly repositioned and reoriented at the end of their lifetime. The density of the ground plane elements was 5.0 elements/m^2. The ground plane was circular, with a radius of 70 m (near and far clipping planes at 5 cm and 40 m, respectively). The animals were re-positioned at the center of this environment at the beginning of each trial. Importantly, this teleportation was undetectable to animals, given that we matched the last frame of trial *t-1* and first of trial *t*. In other words, from the standpoint of the animals, they continuously explored an infinite environment. Spike2 software (Cambridge Electronic Design Ltd., Cambridge, UK) was used to record and store the timeseries of the target and animal's location, animal linear and angular velocity, as well as eye movements. All behavioral data were recorded along with the neural event markers at a sampling rate of 833.33 Hz.

Behavioral task

Monkeys steered to a target location (circular disc of radius 20 cm) that was cued briefly (300 ms) at the beginning of each trial. Each trial was programmed to start after a variable random delay (truncated exponential distribution, range: 0.2 – 2.0 s; mean: 0.5 s) following the end of the previous trial. The targets appeared at a random location between −45 to 45 deg of visual angle, and between 1 and 4 m relative to where the animal was stationed at the beginning of the trial. The joystick was always active, and thus monkeys were free to start moving before the target vanished, or before it appeared. On most trials (88%) the animals were stationary when the target appeared. Monkeys typically performed blocks of 500-750 trials before being given a short break. Feedback was given in the form of juice reward following a variable waiting period (truncated exponential distribution, range: 0.1 – 0.6 s; mean: 0.25 s) after stopping within 0.6 m from the center of the target ("reward boundary"). No juice was provided otherwise. If the animal never started moving after target presentation (i.e., linear velocity was always zero), or did not stop after 7 seconds, the trial would time out and the next inter-trial interval would start.

Neural recording and pre-processing

We recorded neural activity extracellularly, either acutely using a 24-channel linear electrode array (100 μm spacing between electrodes; U-Probe, Plexon Inc, Dallas, TX, USA) or chronically with multi-electrode arrays of either 48 or 96 electrodes (Blackrock Microsystems, Salt Lake City, UT, USA). Parietal area 7a and dorso-medial pre-frontal cortex (dlPFC) were recorded both with linear probes and Utah arrays, while dorso-medial superior temporal area (MSTd), being a deep structure, was only recorded with linear arrays. We confirmed that our results do not depend on type of electrode used (also see Noel et al.). During acute recordings with the linear arrays, the probes were advanced into the cortex through a guide-tube using a hydraulic micro-drive. Spike detection thresholds were manually adjusted separately for each channel to facilitate real-time monitoring of action potential waveforms. Recordings began once waveforms were stable. The broadband signals were amplified and digitized at 20 KHz using a multichannel data acquisition system (Plexon Inc, Dallas, TX, USA). Raw data was stored along with the action potential waveforms for offline analysis. Additionally, for each channel, we also stored low-pass filtered (-3dB at 250 Hz) local-field potentials (LFP). For the array recordings, broadband neural signals were amplified and digitized at 30 KHz using a digital headstage (Cereplex E, Blackrock Microsystems, Salt Lake City, UT, USA), and stored for offline analysis. Additionally, for each channel, we stored low-pass filtered LFPs sampled at 500 Hz. Finally, copies of event markers were received online from the stimulus acquisition software (Spike2) and saved alongside the neural data.

Spike detection and sorting were performed on raw neural signals using KiloSort 2.0 on an NVIDIA Quadro P5000 GPU. The spike clusters produced by KiloSort were visualized in Phy2 and manually refined by a human observer by either accepting or rejecting KiloSort's label for each unit. In addition, we computed three isolation quality metrics; inter-spike interval violations (ISIv), waveform contamination rate (cR), and presence rate (PR). ISIv is the fraction of spikes that occur within 2 ms of the previous spike. cR is the proportion of spikes inside a high-dimensional cluster boundary (by waveform) that are not from the cluster (false positive rate) when setting the cluster boundary at a Mahalanobis distance such that there are equal false positives and false negatives. PR is 1 minus the fraction of 1-minute bins in which there is no spike. We set the following thresholds in qualifying a unit as a single-unit and retaining it for further analyses: ISIv <20%, cR <0.02, and PR > 90%.

Analyses

Behavioral analyses. Trials were classified as "attempted" if they did not satisfy the criteria for the other categories: "never started", "never

stopped", or "aborted". Never started trials were those where upon trial end (trial maximum duration = 7 seconds), the animal had moved less than 5 cm from the origin (typically exactly 0 cm). Never stopped trials were those where the animals' linear velocity was above 5 cm/s at the of the trial (typically 200 cm/s). In these "never stopped" trials, animals typically maintained the joystick at a fixed angle (e.g., 45° upward and leftward), resulting in stereotyped circular trajectories that were disengaged from the task of "catching" fireflies. Aborted trials were those where the animal had started an attempt (linear velocity > 5 cm/s) yet traveled less than 30% of the target distance. For the transition probability matrix in Fig. 1C and run lengths in Fig. 2D rewarded trials were considered "attempted."

To estimate the reward rate afforded by trials (Fig. 1E, G, I), we computed the fraction of attempted trials that were rewarded for the 30 closest trials within the same session. Similarly, in these same trials we computed the mean time it took to reach these targets, hence allowing to estimate a reward rate (i.e., probability of reward divided by time). This estimate was robust to the number of adjacent trials used (range tested: 15-60 trials).

To classify sessions as belonging to different animals, we build a linear discriminant analysis (LDA). The model assumes that each class (Y) generates data (X) given a multivariate normal distribution. The covariance matrix of each class is assumed to be the same (empirically estimated, pooled covariance), with the means varying. We use a 10-fold cross-validation, wherein 90% of the data is used to estimate class boundaries, and 10% is used in testing. We iterate over folds to classify all data. The prediction aims to minimize the expected classification cost:

$$\hat{y} = argmin \sum_{k=1}^{k} \hat{P}(k|x) C(y|k) \tag{5}$$

where $\hat{y}$ is the predicted classification, k is the number of classes, $\hat{P}(k,|,x)$ is the posterior probability of class k for observation x (uniform prior), and $C(y,|,k)$ is the cost of classifying an observation as y when its true class is k. The standard cost for LDA was employed.

To determine to what extent aborting behavior was reward optimizing, we performed a set of behavioral simulations comparing three counterfactual policies: actual, no-aborts, and optimal. In the *actual* policy, we computed reward rates for each session by using each trial's empirical outcome (i.e., rewarded vs. not rewarded) and duration exactly as observed, preserving the animal's natural mixture of attempts and aborts as well as all inter-trial intervals. In the *no-aborts* policy, every aborted trial was re-computed as if the animal had attempted it instead. For each such trial, we estimated the probability of obtaining a reward and the expected trial duration using a k-nearest-neighbors procedure applied within the same session, matching on target distance and bearing. K was set 20 trials. These estimates allowed us to assign a counterfactual reward and time cost to every would-be abort while keeping all other trials unchanged. Finally, in the *optimal* policy, we computed the theoretical upper bound on achievable reward rate. For each trial, we estimated the attempt reward density (probability of reward divided by expected duration, as for the "no-abort" policy) and swept a threshold τ determining whether the agent should attempt or abort the trial. For every τ, we simulated behavior forward in time, including the realistic inter-trial interval, until the total simulated time matched the session's actual duration. The value τ* yielding the highest reward rate defined the optimal policy for that session. Comparing the reward rates across these three policies allowed us to assess whether the animal's empirical pattern of aborting behavior approached, exceeded, or fell short of reward-maximizing performance.

Neural analyses. Spiking activity was digitized in 1 ms bins. Then, within each session (72 in total) each aborted trial was matched to its most similar attempted or aborted trial, by first finding target locations within 5 cm (radial) of the target trial. Then, within this subset (typically a dozen trials), we chose the trial with the smallest mean squared error (MSE) in contrasting linear and angular velocity profiles between target and candidate trials. Animals typically followed curvilinear trajectories (see Fig. 1B) where the angular component was null or close to null during target presentation. Likewise, given that trials are matched within-sessions, matched by target location and velocity also matches reward rates (i.e., likelihood of attaining a particular reward is computed within sessions).

For single neuron analyses, we performed a spike count within 0 to 300 ms from target presentation (i.e., the time-period over which the target was visible), corrected this spike count by trial-specific baseline (-300 to 0 ms), and performed statistical testing comparing matched attempted and aborted trials via a Wilcoxon signed rank test ($p < 0.05$). For visualization, we convolved spike trains with a causal Gaussian kernel (std = 30 ms) and then averaged within conditions. To compare fraction of neurons tuned to aborting behavior across brain regions and to chance, we performed Mann–Whitney U tests. Further, to quantify tuning to task variables (14 in total, see Fig. 4), we fit a fully coupled Poisson Generalized Additive Model (pGAM) to each neuron. The model links spike counts $y_t \in \mathbb{N}_0$ to continuous covariates x_t (linear and angular velocity and acceleration, linear and angular distance traveled, linear and angular distance to target) and discrete covariates z_t (movement onset/offset, target onset, reward delivery, and simultaneously recorded population activity). The log firing rate is expressed as

$$\log \mu = \sum_j f_j\left(x_j\right) + \sum_k f_k{}^* z_k \tag{6}$$

where * is the convolution operator, and the unit spike counts are generated as Poisson random variables with rate specified by Eq. 6. Each nonlinearity $f(\cdot)$ is parameterized by B-splines with a roughness penalty,

$$PEN\left(f,\lambda_f\right) = -\frac{1}{2}\lambda_f \boldsymbol{\beta}^T S_f \boldsymbol{\beta},\ S_f = \int b'' \cdot b''^T \mathrm{d}x \tag{7}$$

The larger λ_f, the smoother the model. This penalization terms can be interpreted as (improper) Gaussian priors over model parameters, the resulting log-likelihood of the model takes the form,

$$\mathscr{L}(\mathbf{y}) = \log p(\boldsymbol{y},|,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\beta}) + \sum_f PEN(f,\lambda_f) \tag{8}$$

The probabilistic form of the penalty enables computation of posterior parameter distributions and confidence intervals with appropriate frequentist coverage, allowing statistical testing for the inclusion of each variable. Model fits were retained only if the pseudo-R^2 on held-out data (20%) exceeded 0.05, and a neuron was considered tuned to a variable when the variable-inclusion test reached $p < 0.01$.

For population-level analysis, we again matched pairs of attempted-aborted or aborted-aborted trials, and expressed the neural activity (averaged between 0 and 300 ms post-stimulus onset and corrected for baseline) of n neurons in an n-dimensional space (Fig. 3F). We then computed the Euclidian distance between aborted and attempted trials (d_1), as well as two aborted trials (d_2). We averaged these distances within sessions, and plot histograms of average d_1 and d_2 distances across sessions (i.e., Fig. 5A, each count is a session). To quantify the difference in distributions we employ KL divergence. For PCA (Fig. 5C and Fig. S6), we convolved spike trains (as above), then averaged and finally performed PCA. We visualize the first PC, accounting for ~69% of the total variance, and perform statistical testing at each time (from -100 to 1300 ms post-target onset) point via Wilcoxon signed rank test. For the dPCA analyses spike trains were

smoothed (Gaussian kernel, σ = 30 ms), baseline-corrected (−200–0 ms), and arranged per session into a tensor R(neurons × time × condition × matched-pairs), where conditions were attempted vs. aborted trials. We used the standard dPCA implementation[38] with marginalizations over condition and time. Regularization was chosen via within-session cross-validation (leave-one-pair-out), and components explaining <2% of variance were discarded. We summed over all marginalization for a given factor to compute total variance explained by different subspaces. For visualization, held-out data were projected onto the first condition or time dPC, and attempted vs. aborted trajectories were compared over −200 to 1000 ms using a Wilcoxon signed-rank test at each time point.

Actor-critic RL models

The actor-critic models used in the current report are augmented version from those presented in Zhang et al., 2024[22]. The formal detail regarding definition of state, action, observation, reward and state transitions are presented here in the *Supplementary Information*. Similarly, formulas driving the actor and critic updates, and belief modeling, as well as detail to the neural architectures and agent training can be found in the Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data generated in this study have been deposited here: https://osf.io/eybc8/. Source data are provided with this paper.

Code availability

Code for reinforcement learning models (including use instructions) is available here: (https://github.com/ryzhang1/dlPFC-strategic-aborting).

References

1. von Neumann J., Morgenstern O. *Theory of Games and Economic Behavior*. 60th Anniversary Commemorative Edition. Princeton University Press. (1944).
2. Parker, G. A. & Smith, J. M. Optimality theory in evolutionary biology. *Nature* **348**, 27–33 (1990).
3. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, USA, 1998).
4. Daw, N. D. & Tobler, P. N. Value learning through reinforcement: the basics of dopamine and reinforcement learning. Neuroeconomics: Decision Making and the Brain 2nd edn. (eds. Glimcher, P. W. & Fehr, E.) 283–298 (Academic, New York, 2014).
5. Wang, J. X. et al. Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
6. Rushworth, M. F. & Behrens, T. E. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **11**, 389–397 (2008).
7. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
8. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
9. Botvinick, M. et al. Reinforcement Learning, Fast and Slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
10. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).
11. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nat. Sep* **11**, 227–231 (2008).
12. Middlebrooks, P. G. & Sommer, M. A. Neuronal correlates of metacognition in primate frontal cortex. *Neuron* **75**, 517–530 (2012).
13. Steiner, A. P. & Redish, A. D. Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. *Nat. Neurosci.* **17**, 995–1002 (2014).
14. Kennerley, S. W., Dahmubed, A. F., Lara, A. H. & Wallis, J. D. Neurons in the frontal lobe encode the value of multiple decision variables. *J. Cogn. Neurosci.* **21**, 1162–1178 (2009).
15. Genovesio A., Brasted P. J., Mitz A. R., Wise S. P. Prefrontal cortex activity related to abstract response strategies. *Neuron*. **47**, 307–320 (2005).
16. Matsuzaka, Y., Akiyama, T., Tanji, J. & Mushiake, H. Neuronal activity in the primate dorsomedial prefrontal cortex contributes to strategic selection of response tactics. *Proc. Natl. Acad. Sci. USA* **109**, 4633–4638 (2012).
17. Shallice, T. & Burgess, P. W. Deficits in strategy application following frontal lobe damage in man. *Brain* **114**, 727–741 (1991).
18. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
19. Lakshminarasimhan K. J. et al. A Dynamic Bayesian observer model reveals origins of bias in visual path integration. *Neuron*. **99**, 194–206 (2018).
20. Lakshminarasimhan K. J. et al. Tracking the mind's eye: primate gaze behavior during virtual visuomotor navigation reflects belief dynamics. *Neuron*. **106**, 662–674 (2020).
21. Noel, J. P. et al. Supporting generalization in non-human primate behavior by tapping into structural knowledge: Examples from sensorimotor mappings, inference, and decision-making. *Prog. Neurobiol.* **201**, 101996 (2021).
22. Zhang, R., Pitkow, X. & Angelaki, D. E. Inductive biases of neural network modularity in spatial navigation. *Sci. Adv.* **10**, eadk1256 (2024).
23. Driscoll L., Shenoy K., Sussillo D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nat Neurosci.* **27**, 1349–1363 (2024).
24. Noel, J. P. et al. Coding of latent variables in sensory, parietal, and frontal cortices during closed-loop virtual navigation. *Elife* **11**, e80280 (2022).
25. Noel, J. P., Balzani, E., Savin, C. & Angelaki, D. E. Context-invariant beliefs are supported by dynamic reconfiguration of single unit functional connectivity in prefrontal cortex of male macaques. *Nat. Commun.* **15**, 5738 (2024).
26. Lakshminarasimhan K. J., Avila E., Pitkow X., Angelaki D. E. Dynamical latent state computation in the male macaque posterior parietal cortex. *Nat. Commun.* **14**, 1832 (2023).
27. Ashwood Z. C. et al. Mice alternate between discrete strategies during perceptual decision-making. *Nat. Neurosci.* **25**, 201–212 (2022).
28. Bruijns, S. A. et al. (2023). Dissecting the complexities of learning with infinite hidden Markov models. bioRxiv, 2023-12
29. Fisher, E. M. Linear Discriminant Analysis. *Stat. Discret. Methods Data Sci.* **392**, 1–5 (1936).
30. Åström, K. J. Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.* **10**, 174–205 (1965).
31. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134 (1998).
32. Konda, V., & Tsitsiklis, J. (1999). Actor-critic algorithms. In Advances in Neural Information Processing Systems 12
33. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

34. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
35. Levine, S., Kumar, A., Tucker, G., & Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. Preprint at arXiv. https://arxiv.org/abs/2005.01643 (2020).
36. Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020).
37. Park, S. Frans, K., Levine, S., Kumar. A. Is value learning really the main bottleneck in offline RL? *arXiv* preprint arXiv:2406.09329 (2024).
38. Kobak D. et al. Demixed principal component analysis of neural population data. *Elife*. **5**, e10989. (2016).
39. Alefantis P. et al. Sensory evidence accumulation using optic flow in a naturalistic navigation task. *J. Neurosci.* Epub **42**, 5451–5462 (2022).
40. Lak, A. et al. Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201 (2014).
41. Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., Kepecs, A. Striatal dopamine mediates hallucination-like perception in mice. *Science*. **372**, eabf4740 (2021).
42. Mah, A., Schiereck, S. S., Bossio, V., Constantinople, C. M. Distinct value computations support rapid sequential decisions. *Nat Commun*. **14**, 7573 (2023).
43. Mah, A., Golden, C., Constantinople, C. M. Dopamine transients encode reward prediction errors independent of learning rates. *Cell Rep.* **43**, 114840 (2024).
44. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
45. Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T. & Miyamoto, A. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat. Neurosci.* **16**, 749–755 (2013).
46. Zylberberg, A., Fetsch C. R., Shadlen M. N. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *Elife*. **5**, e17688 (2016).
47. Medendorp, W. P., Heed, T. State estimation in posterior parietal cortex: distinct poles of environmental and bodily states. *Prog Neurobiol.* **183**, 101691 (2019).
48. Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E. & Behrens, T. E. Frontal cortex and reward-guided learning and decision-making. *Neuron* **70**, 1054–1069 (2011).
49. Rudebeck, P. H. et al. Frontal cortex subregions play distinct roles in choices between actions and stimuli. *J. Neurosci.* **28**, 13775–13785 (2008).
50. Averbeck B., O'Doherty J. P. Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*. **47**, 147–162 (2021).
51. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340 (2005).
52. Seo, M., Lee, E. & Averbeck, B. B. Action selection and action value in frontal-striatal circuits. *Neuron* **74**, 947–960 (2012).
53. Nonomura, S. et al. Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. *Neuron* **99**, 1302–14.e5 (2018).
54. Yttri, E. A. & Dudman, J. T. Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature* **533**, 402–406 (2016).
55. Collins, A. G. & Frank, M. J. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* **121**, 337 (2014).
56. Soltani A., Koechlin E. Computational models of adaptive behavior and prefrontal cortex. *Neuropsychopharmacology*. **47**, 58–71 (2022).
57. Fujii, N. & Graybiel, A. M. Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science* **301**, 1246–1249 (2003).
58. Averbeck, B. runoB. & Lee, D. aeyeol Prefrontal neural correlates of memory for sequences. *J. Neurosci.* **27**, 2204–2211 (2007).
59. Averbeck, B. runoB. et al. Parallel processing of serial movements in prefrontal cortex. *Proc. Natl. Acad. Sci.* **99**, 13172–13177 (2002).
60. Cai Y. et al. Time-sensitive prefrontal involvement in associating confidence with task performance illustrates metacognitive introspection in monkeys. *Commun. Biol.* **5**, 799 (2022).
61. Wang S., Falcone R., Richmond B., Averbeck B. B. Attractor dynamics reflect decision confidence in macaque prefrontal cortex. *Nat. Neurosci*. **26**, 1970–1980 (2023).
62. Gao, Y., Xue, K., Odegaard, B. & Rahnev, D. Automatic multisensory integration follows subjective confidence rather than objective performance. *Commun. Psychol.* **3**, 38 (2025).
63. Odegaard B. et al. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proc. Natl. Acad. Sci. USA*. **115**, E1588–E1597 (2018).
64. Kubota, K., Komatsu, H. Neuron activities of monkey prefrontal cortex during the learning of visual discrimination tasks with GO/NO-GO performances. *Neurosci. Res*. **3**, 106–129 (1985).
65. Sakagami, M., Niki, H. Spatial selectivity of go/no-go neurons in monkey prefrontal cortex. *Exp Brain Res*. **100**, 165–169 (1994).
66. Sakagami, M. & Tsutsui, K. The hierarchical organization of decision making in the primate prefrontal cortex. *Neurosci. Res* **34**, 79–89 (1999). PMID: 10498334.
67. van Gaal, S., Ridderinkhof, K. R., Fahrenfort, J. J., Scholte, H. S., Lamme, V. A. Frontal cortex mediates unconsciously triggered inhibitory control. *J. Neurosci.* **28**, 8053–8062 (2008).
68. Bruni, S., Giorgetti, V., Bonini, L., Fogassi, L. Processing and integration of contextual information in monkey ventrolateral prefrontal neurons during selection and execution of goal-directed manipulative actions. *J. Neurosci*. **35**, 11877–90 (2015).
69. Johnston, W. J., Fine, J. M., Yoo, S. B. M., Ebitz, R. B., Hayden, B. Y. Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. *Nat. Neurosci.* **27**, 2218–2230 (2024).
70. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature*. **497**, 585–590 (2013).
71. Bernardi, S. et al. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954–967 (2020).
72. Kwon, M., Daptardar, S., Schrater, P. R. & Pitkow, X. Inverse rational control with partially observable continuous nonlinear dynamics. *Adv. Neural Inf. Process Syst.* **33**, 7898–7909 (2020).

Acknowledgements

The authors thank Jing Lin and Jian Chen for programming the experimental stimulus. We also thank Eric Avila, Stefania Bruni, and Panos Alefantis for data collection; Roozbeh Kiani for surgical expertise; and Kaushik Lakshminarasimhan for helpful discussions leading to the original RL agents.

Author contributions

Conceptualization: J.P.N., R.Z., X.P., D.A.; Data Curation: J.P.N., R.Z.; Formal Analysis: J.P.N., R.Z.; Funding Acquisition: J.P.N., D.A.; Investigation: J.P.N., R.Z.; Methodology: J.P.N., R.Z.; Project Administration: J.P.N., X.P., D.A.; Resources: R.Z., X.P.; Software: R.Z., X.P.; Supervision: X.P., D. A.; Validation: J.P.N., R.Z., X.P., D.A.; Visualization: J.P.N., R.Z.; Writing – original draft: J.P.N., R.Z.; Writing – review & editing: J.P.N., R.Z., X.P., D.A. D.A. and X.P., jointly supervised this work.

Funding

The work was supported by 1U19NS118246, 1R01NS120407, and 1R01DC014678 from NIH to D.E.A. X. P., D.E.A., and J.P.N were supported by SFI-AN-NC-SCN-00007276. J.P.N was additionally supported by NIH

R00NS128075, SFI-AN-AR-Pilot-00008952, CTSI 1UM1TR004405, and a Sloan Research Fellowship.

Competing interests
The authors declare no competing interests, financial or other.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-74783-6.

Correspondence and requests for materials should be addressed to Jean-Paul Noel.

Peer review information *Nature Communications* thanks Yoshiya Matsuzaka, Hua Tang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.