# Deliberation gated by opportunity cost adapts to context with urgency

Maximilian Puelma Touzel,[1,2,*] Paul Cisek,[3] and Guillaume Lajoie[1,4,5]

[1] *Mila, Quebec AI Institute*

[2] *Department of Computer Science and Operations Research, Université de Montréal*

[3] *Department of Neuroscience, Université de Montréal*

[4] *Department of Mathematics and Statistics, Université de Montréal*

[5] *Canada CIFAR AI Chair*

## Abstract

The value we place on our time impacts what we decide to do with it. Value it too little, and we obsess over all details. Value it too much, and we rush carelessly to move on. How to strike this often context-specific balance is a challenging decision-making problem. Average-reward, putatively encoded by tonic dopamine, serves in existing reinforcement learning theory as the stationary opportunity cost of time. However, environmental context and the cost of deliberation therein often varies in time and is hard to infer and predict. Here, we define a non-stationary opportunity cost of deliberation arising from performance variation on multiple timescales. Estimated from reward history, this cost readily adapts to reward-relevant changes in context and suggests a generalization of average-reward reinforcement learning (AR-RL) to account for non-stationary contextual factors. We use this deliberation cost in a simple decision-making heuristic called *Performance-Gated Deliberation*, which approximates AR-RL and is consistent with empirical results in both cognitive and systems decision-making neuroscience. We propose that deliberation cost is implemented directly as urgency, a previously characterized neural signal effectively controlling the speed of the decision-making process. We use behaviour and neural recordings from non-human primates in a non-stationary random walk prediction task to support our results. We make readily testable predictions for both neural activity and behaviour and discuss how this proposal can facilitate future work in cognitive and systems neuroscience of reward-driven behaviour.

Keywords: primate decision-making, reinforcement learning, urgency, opportunity cost

* puelmatm@mila.quebec

| symbol | quantity |
|---|---|
| $t$ | within-trial time |
| $k$ | trial index |
| $S_t$ | within-trial state at time $t$ |
| $\boldsymbol{S}_t$ | state sequence up to time $t$ |
| $R_k$ | reward of $k$th trial |
| $T_k$ | duration of $k$th trial |
| $t_k^{\text{dec}}$ | decision time of $k$th trial |
| $\mathcal{C}_t^{\text{del}}$ | within-trial opportunity cost of deliberation |
| $r_{\max}$ | maximum reward acheiveable in a trial |
| $b_t$ | belief of correct report given $\boldsymbol{S}_t$ |
| $\bar{r}_t$ | expected reward for reporting at time $t$ |
| $\mathcal{C}_t^{\text{com}}$ | within-trial opportunity cost of commitment |
| $\rho$ | stationary reward rate |
| $\rho^*$ | optimal stationary reward rate |
| $\alpha$ | context parameter |
| $\rho_\alpha$ | context-conditioned stationary reward rate |
| $T_\alpha$ | context-conditioned stationary average trial duration |
| $\hat{\rho}_k^\tau$ | reward history filtered through a timescale, $\tau$ |
| $\tau_{\text{long}}$ | a long timescale over which to estimate $\rho$ |
| $\tau_{\text{context}}$ | a context-specific timescale over which to estimate $\rho_\alpha$ |
| $\nu$ | tracking cost sensitivity |
| $K$ | subjective reward scale factor |
| $T_{\text{block}}$ | characteristic duration of a trial block |
| $c$ | auxiliary deliberation cost rate |
| $N_t$ | tokens difference |
| $p$ | jump probability of random walk, $p \geq 1/2$ |

Table I. Symbol glossary. Highlighted in gray are parameters of the PGD model presented in this paper.

## INTRODUCTION

Humans and other animals make a wide range of decisions throughout their daily lives. Any particular action usually arises out of a hierarchy of decisions involving a careful balance between resources, including one that is always limited: time. The cost of *spending* time depends on its value, a construct that relies on comparing against the alternative things an agent could potentially do with it. Estimating time's value is not straightforward for a number of reasons. There are alternative choices at multiple decision levels, e.g. moving on from a job and moving on from a career, and each level requires its own evaluation. Moreover, the value of alternatives needs to be tracked as they may change over time depending on the context in which a decision is made. For example, animals will learn to value a given food resource differently depending on whether it is encountered during times of plenty versus scarcity. The agent's knowledge of and ability to track context thus influences the value it assigns to possible alternatives.

These are significant, practical complications of making decisions contingent on *opportu-*

*nity costs* [1], the formal economic concept capturing the value of the alternatives lost by committing a limited resource to a given use. The opportunity cost of time is nevertheless well-studied in decision-making theory for relative definitions of value, most notably as the average reward in average-reward reinforcement learning (AR-RL) [2]. AR-RL focuses on deviations from the average reward rather than on discounted reward as in the more widely known discount-reward reinforcement learning formulation. In neuroscience, AR-RL was first proposed to extend the reward prediction error hypothesis for phasic dopamine to account also for the observed properties of tonic dopamine levels [3]. It has since been used to emphasize the relative nature of reward-based decision-making [4] in explanations of human and animal behaviour in foraging [5], free-operant conditioning [6], perceptual decision-making [7, 8], cognitive effort/control [8, 9], and even economic exchange [10].

AR-RL is increasingly acknowledged as the more suitable reinforcement learning formulation [11] for *continuing environments* in which there is no definite end [12], in a large part because it explicitly seeks solutions that achieve the maximum possible average reward rate. This is in contrast to traditional fixed accuracy criteria in perceptual decision-making tasks that focus on maximizing trial reward alone [13]. The solutions to AR-RL formulations of tasks of long sequence of trials are decision boundaries in the state space of a trial that typically collapse with trial time. This limits deliberation in trials with low return-on-time-investment, e.g. in difficult trials for tasks in which trial difficulty is variable [7, 14].

Up to now, however, AR-RL and most of its applications have focused on fixed context and have used the stationary average reward as the fixed opportunity cost of time, which ignores context-dependent performance variation. This is perhaps not surprising given that in psychological and neuroscientific studies of decision-making, we usually eliminate such contextual factors from the experimental design such that our models describe stationary behaviour. However, the brain mechanisms under study are adapted to a more diverse natural world in which contextual factors are often relevant, hard to infer and vary over time [4].

We pursue a theory of approximate relative-value decision-making under uncertainty in a setting relevant to decision-making neuroscience. We start by showing that value in AR-RL can be expressed using the opportunity costs of deliberation and commitment. Here, the commitment cost is the shortfall in reward relative to the maximum possible in a trial that is expected to be lost when committing to a decision at a given time. Highlighting the risk of value representations in non-stationary environments, we propose an approximation to the AR-RL value-optimal solution, Performance-Gated Deliberation (PGD), that uses the opportunity cost directly as the collapsing decision boundary, instead of as input to a value optimization problem. PGD thus reduces decision-making to estimating two opportunity costs: a commitment cost learned from the statistics of the environment and a deliberation cost estimated from tracking one's own performance in that environment. It explains how an agent, without explicitly tracking context parameters or storing a value function, can trade-off speed and accuracy according to performance at the typically longer timescales over which context changes. We propose that deliberation cost is then directly encoded as "urgency" in the neural dynamics underlying decision-making [7, 15–17]. The theory is thus directly testable using both behaviour and neural recordings. To illustrate how PGD applies in a specific continuing decision-making task, and to make the links to a neural implementation explicit, we analyze behavior and neural recordings collected over eight years from two non-human primates (NHPs) [18, 19]. They performed successive trials of the "tokens task", a probabilistic guessing task in which information about the correct choice is continuously changing within each trial, and a task parameter controlling the incentive to decide early (the context) is varied over longer timescales. Behavior in the task, in both humans [16] and monkeys [19], provides additional support to an existing hypothesis about how neural dynamics implements time-sensitive decision-making [15]. Specifically, neural recordings in monkeys suggest that the evidence needed to make the decision predominates in dorsolateral prefrontal cortex [20]; a growing context-dependent urgency signal is provided by the basal ganglia [21]; and the two are combined to bias and time, respectively, a competition between potential actions that unfolds in dorsal premotor and primary motor cortex [18]. Similar findings have been reported in other tasks - for example, in the frontal eye fields during decisions about eye-movements [17]. PGD is proposed as the theoretical explanation for why decision-making mechanisms are organized in this way. As an algorithm, it serves as a robust means to balance immediate rewards and the cost of time across multiple timescales. As a quantitative model, it serves to explain concurrently recorded behaviour and neural urgency in continuing decision-making tasks.

## RESULTS

### A. Theory of performance-gated deliberation

#### 1. *Opportunity costs of deliberation and commitment, and drawbacks of average-reward reinforcement learning*

We consider a class of tasks consisting of a long sequence of trials indexed by $k = 1, 2, \dots$ (see fig. 1a), each of which provides the opportunity to obtain some reward by choosing correctly. In each trial, a finite sequence of states, $S_t$, $t = 0, \dots, t_{\max}$, is observed that provide evidence for an evolving belief about the correct choice among a fixed set of options. To keep notation simple, we suppress denoting the trial index, $k$, on quantities such as trial state, $S_t$, that also depend on trial time, $t$. The time of decision, $t_k^{\mathrm{dec}}$, and the chosen option determine both the reward received, $R_k$, and the trial duration, $T_k \geq t_k^{\mathrm{dec}}$. Importantly, decision timing can affect performance because earlier decisions typically lead to shorter trials (and thus more trials in a given time window), while later decisions lead to higher accuracy. Effectively balancing such speed-accuracy trade-offs is central to performing well in continuing episodic task settings. For a fixed strategy, the *stationary reward rate* (see slope of dashed line in fig. 1a(right)) is

$$\rho := \lim_{k\to\infty} \sum_k R_k \Big/ \sum_k T_k \, . \tag{1}$$

For a stochastic environment, the definition of $\rho$ includes an ensemble average. Free-operant conditioning, foraging, and several perceptual decision-making tasks often fall into this class. Previous work [7, 22] has studied the belief of correct report for binary rewards, $b_t = P(R_k = 1|\boldsymbol{S}_t, t^{\mathrm{dec}} = t)$, which also gives the expected trial reward, $\bar{r}_t = b_t \cdot 1 + (1 - b_t) \cdot 0 = b_t$ [7] (see [23] for more about the relationship between value-based and perceptual decisions). $\boldsymbol{S}_t$ denotes the state sequence observed so far, $(S_0, \dots, S_t)$. We consider greedy strategies that report the choice with the largest belief at decision time. The decision problem is then about *when* to decide.

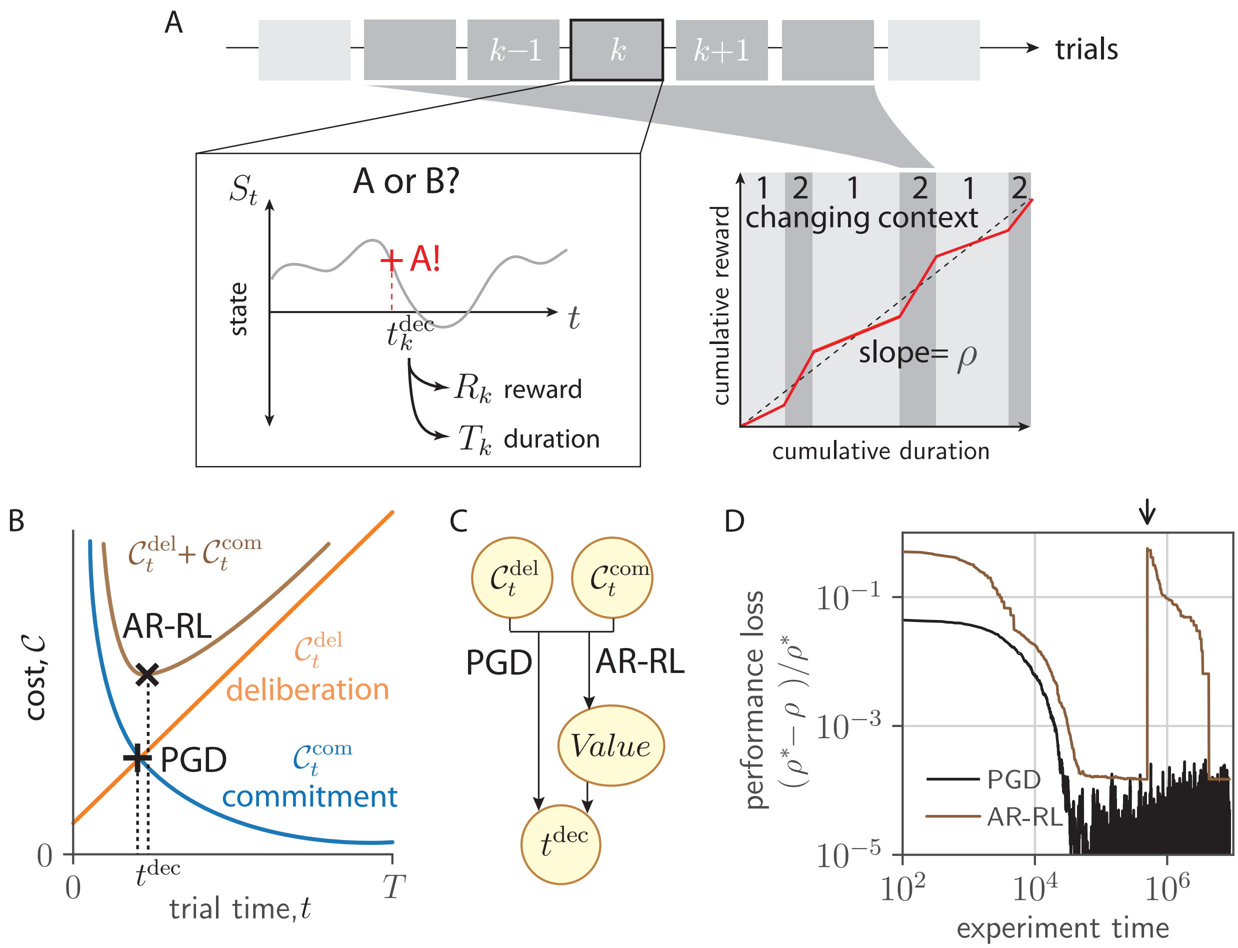



Figure 1. *AR-RL and Performance-Gated Deliberation.* (a) Task setting. Left: Within trial state, $S_t$ evolves over trial time $t$ in successive trials indexed by $k$. The decision 'A' is reported at the decision time $t_k^{\mathrm{dec}}$ (red cross), determining trial reward, $R_k$, and trial duration, $T_k$. Right: Sketch of cumulative reward versus cumulative duration. Context-conditioned reward rate (slope of red line), varies with alternating context (labelled 1 and 2) around average reward, $\rho$ (dashed line). (b) Decision rules based on opportunity costs of commitment, $\mathcal{C}_t^{\mathrm{com}}$, and deliberation, $\mathcal{C}_t^{\mathrm{del}}$. The AR-RL rule (black 'x') finds $t$ that minimizes $\mathcal{C}_t^{\mathrm{del}} + \mathcal{C}_t^{\mathrm{com}}$. The PGD rule (black cross) finds $t^{\mathrm{dec}}$ at which they intersect, $\mathcal{C}_t^{\mathrm{del}} = \mathcal{C}_t^{\mathrm{com}}$. (c) Schematic diagram of each algorithm's dependency. PGD computes a decision time directly from the two opportunity costs, while AR-RL uses both to first estimate a value function, whose maximum specifies the decision time. (d) Loss (error in performance with respect to the optimal policy, $(\rho^* - \rho)/\rho^*$) over learning time in a patch-leaving task (AR-RL: brown, PGD: black). The arrow indicates when the state labels were randomly permuted.

Average-reward reinforcement learning (AR-RL), first proposed in artificial intelligence [24], was later incorporated into reward prediction error theories of dopamine signalling [3] and employed to account for the opportunity cost of time [6]. AR-RL was subsequently used to study reward-based decision-making in neuroscience and psychology [7, 8, 25, 26]. AR-RL centers around the average-adjusted future return, which penalizes the passage of time according the average reward. A reporting decision is associated with a return that for trial-based tasks combines the remainder of the current trial and all future trials, $\bar{r}_t - \rho(T_k - t) + \sum_{k'>k}(R_{k'} - \rho T_{k'})$, where $\rho$ (*c.f.* eq. (1)) is either estimated online or obtained self-consistently (see Methods for details). Value is defined as the future return averaged over trial sequence realizations. This average of a sum of reward deviations into the future converges on account of the decaying effects of the state at which the decision is made. The AR-RL algorithms we consider aim to achieve the highest $\rho$ by also maximizing the average-adjusted value. We now provide an alternative, but equivalent definition of average-adjusted trial return in terms of opportunity costs incurred by the agent.

We denote the opportunity cost of committing at time $t$ within a trial as $\mathcal{C}_t^{\text{com}}$, defined as the difference

$$\mathcal{C}_t^{\text{com}} = r_{\max} - \bar{r}_t \;, \tag{2}$$

where $r_{\max}$ is the maximum trial reward possible *a priori*. Within a trial, an agent lowers its commitment cost towards zero by accumulating more evidence, i.e. by waiting. Waiting, however, incurs another opportunity cost: the reward lost by not acting. We denote this opportunity cost of deliberation incurred up to a time $t$ in a trial as $\mathcal{C}_t^{\text{del}}$. In AR-RL, the constant opportunity cost rate of time is integrated so that for $T_k = t_k^{\text{dec}}$,

$$\mathcal{C}_t^{\text{del}} = \rho t \;. \tag{3}$$

With these definitions, the average-adjusted trial return for deciding at a time $t$ can be expressed as $r_{\max} - (\mathcal{C}_t^{\text{com}} + \mathcal{C}_t^{\text{del}})$. It is maximized by jointly minimizing $\mathcal{C}_t^{\text{del}}$ and $\mathcal{C}_t^{\text{com}}$ (fig. 1b), giving the AR-RL optimal solution (see Methods for a formal statement and solution of the AR-RL problem). Expressed in this way, the average-adjusted trial return emphasizes the more general perspective that an agent's solution to the speed-accuracy trade-off is about how it balances the decaying opportunity cost of commitment and the growing opportunity cost of deliberation.

Despite their utility, value representations such as the average-adjusted trial return can be a liability in real world tasks where task statistics are non-stationary. To illustrate this, we consider the following foraging task. An foraging agent feeds among a fixed set of food (e.g. berry) patches. Total berries consumed in a patch saturates with duration $t$ according to a given saturation profile, shared across patches, as the fewer berries left are harder to find. Patches differ in their richness (e.g. berry density), which is randomly sampled and fixed over the task. Denoting patch identity (serving as context) by $s$, the food return is directly observed and deterministic given $s$. To perform well, the agent needs to decide when to move on from depleting the current patch. Further details about the task and its solution are given in the Methods. For a broad class of online AR-RL algorithms, the agent learns the average-adjusted trial return as a function of state and time. For a given patch, it then leaves when this return is at its maximum (*c.f.* fig. 1b). In fig. 1d, we show how the performance (brown line) approaches that of the optimal policy in time as the estimation of the AR-RL trial return improves with experience (see Methods for implementation details). However, if the agent's environment undergoes a significant disturbance (e.g. a forest fire due to which the patch locations are effectively re-sampled), the performance of this AR-RL algorithm can drop back to where it started. We implement such a disturbance via random permutation of the state labels at the time indicated by the arrow in fig. 1d. This is true over a range of learning rates and the number of patches (fig. S8). More generally, any approach that relies on estimating state-value associations shares this drawback, including those approaches that implicitly learn those associations by directly learning a policy instead [27]. Could context-dependent decision times be obtained without having to associate value or action to state? A means to do so is presented in the next section.

## 2. *Performance-Gated Deliberation*

We propose that instead of maximizing value as in AR-RL, which minimizes the sum of the two opportunity costs, $\mathcal{C}_t^{\text{del}} + \mathcal{C}_t^{\text{com}}$, the agent simply takes as its decision criterion when they intersect (shown as the black cross in fig. 1b).

$$t^{\text{dec}} := \min_t \left\{ t \mid \mathcal{C}_t^{\text{del}} \geq \mathcal{C}_t^{\text{com}} \right\} \qquad \text{(PGD decision rule)} \tag{4}$$

We call this heuristic rule at the center of our results *Performance-Gated Deliberation* (PGD). Plotted alongside the AR-RL performance in fig. 1d for our example foraging task, PGD (black line) achieves better performance than AR-RL overall. It is also insensitive to the applied disturbance since PGD uses $\mathcal{C}_t^{\text{del}}$ and $\mathcal{C}_t^{\text{com}}$ directly when deciding, rather than as input to problem of optimizing average-adjusted value as in AR-RL (fig. 1c).

We constructed the above task so that PGD is the AR-RL optimal solution. In general, however, PGD is a well-motivated approximation to the optimal strategy, so we call it a heuristic. In the more general stochastic setting where there is residual uncertainty in trial reward at decision time, the PGD agent will have to learn the association between state and expected reward, $\bar{r}_t$. This association is learned from within-trial correlations only. In contrast, the opportunity cost of time as the basis for the deliberation cost depends on across-trial correlations that together determine the overall performance. It is thus more susceptible to non-stationarity. A typical task setting is when the value of the same low-level action plan differs across context. From hereon, we will assume the agent has learned the stationary opportunity cost of commitment and so focus on resolving the remaining problem: how to learn and use an opportunity cost of deliberation that exhibits non-stationarity on the longer timescales over which context varies.

## 3. *Reward filtering for a dynamic opportunity cost of deliberation*

The state disturbance in the toy example above altered task statistics at only a single time point. In general, however, changes in task statistics over time can occur throughout the task experience. A broader notion of deliberation cost beyond the static average reward is thus needed–one that can account for extended timescales over which performance varies. Such a cost serves as a dynamic reference in a relative definition of value based on a non-stationary opportunity cost of time. We first address how performance on various timescales can be estimated.

As a concrete example, we make use of the task that we will present in detail in the following section. This task has a context parameter, $\alpha$, that can vary in time on characteristic timescales longer than the moment-to-moment and can serve as a source of non-stationarity in performance. Here, the context sequence, $\boldsymbol{\alpha}_k$, varies on a single timescale, e.g. through periodic switching between two values. The resulting performance (fig. 2a(top)) varies around the stationary average, $\rho$ (purple), with context variation due to the switching (orange), as well as context-conditioned trial-to-trial variation (blue). The decomposition of time-varying performance into these multiple, timescale-specific components can be achieved by passing the reward signal through parallel filters, each designed to retain the signal variation specific to that timescale (fig. 2a(bottom)). There are multiple approaches to this decomposition. We chose a heuristic approach in which the performance over a finite memory timescale can be estimated by filtering the sequence of rewards through a simple low-pass filter [8, 28].

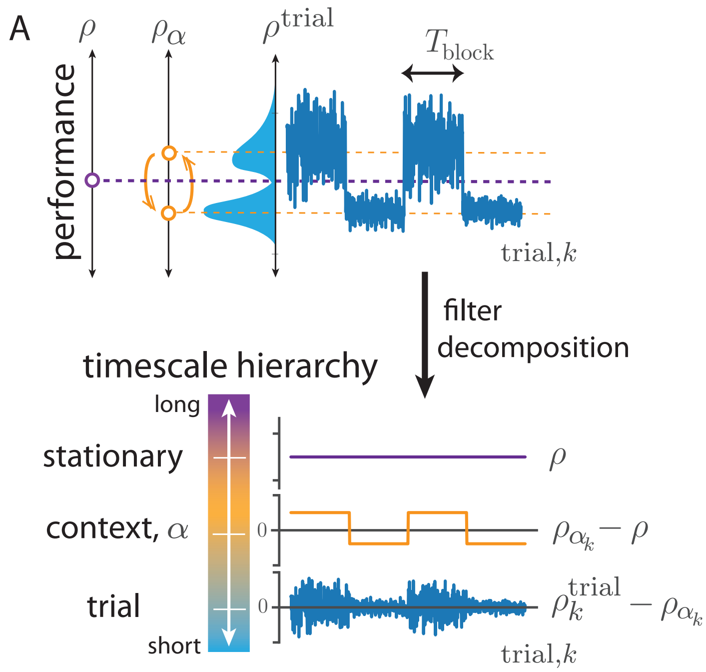



Figure 2. *Non-stationary opportunity cost.* (a) Top: Dynamics of trial performance ($\rho_k^{\text{trial}} := R_k/T_k$; blue) with its distribution as well as dynamics of between context-conditioned averages of performance ($\rho_\alpha = \langle \rho_k^{\text{trial}} \rangle_{k|\alpha}$; orange), and the effectively stationary average performance ($\rho \sim \langle \rho_k^{\text{trial}} \rangle_k$; purple). Bottom: these are decomposed into a hierarchy by filtering reward history on trial, context, and long timescales, respectively. (b) Two hypothetical forms for context-specific trial opportunity cost. Top: Trial-unaware cost in which context varies the slope around $\rho$. Bottom: Trial-aware cost in which context variation is through a bias ( eq. (5)).

This filter is defined by an integration time, $\tau$, tuned to trade off the bias and variance of the estimate in order to best capture the variation on the desired timescale (e.g. how performance varies over different contexts). We denote such an estimate $\hat{\rho}_k^{\tau}$, and show in the Methods that it approximates the average reward over the last $\tau$ time units. We discuss the question of biological implementation in the discussion, but note here that the number and values of $\tau$ needed to represent performance variation in a given task could be learned or selected from a more complete set in an online fashion during task learning. In an experimental setting, these learned values can in principle be inferred from observed behaviour and we developed such an approach in the analysis of data that we present in the following section.

Applying this heuristic decomposition here, the stationary reward rate, $\rho$, can be estimated to high precision by using a long integration time, $\tau_{\text{long}}$, to the reward sequence $R_k$, producing the estimate $\hat{\rho}_k^{\tau_{\text{long}}}$. If $\boldsymbol{\alpha}_k$ were a constant sequence, $\mathcal{C}_t^{\text{del}} = \hat{\rho}_k^{\tau_{\text{long}}} t$, the stationary opportunity cost of deliberation eq. (3) of AR-RL. However, in this example context varies on a specific timescale, to which the former is insensitive. Thus, a second filtered estimate $\hat{\rho}_k^{\tau_{\text{context}}}$ is needed to estimate performance on this timescale. Unlike $\hat{\rho}_k^{\tau_{\text{long}}}$, this estimate tracks the effective instantaneous, context-specific performance, $\rho_{\alpha_k}$. Its estimation error arises from a trade-off, controlled by the integration time, $\tau_{\text{context}}$, between its speed of adaptation and its finite memory.

We consider two distinct hypotheses for how to extend AR-RL to settings where performance varies over context. The first hypothesis, $\mathcal{C}_t^{\mathrm{del}} = \rho_\alpha t$, is the straightforward, *trial-unaware* extension of eq. (3), shown in fig. 2b(top). Here, performance is tracked only on a timescale sufficient to capture context variation and the corresponding cost estimate, $\hat{\rho}_{k-1}^{\tau_{\mathrm{context}}}$, is incurred moment-to-moment, neglecting the trial-based task structure. However, this incorrectly lumps together two distinct opportunity costs: those incurred by moment-by-moment decisions and those incurred as a result of the effective planning implied by performance that varies over context. In particular, context is defined over trials not moments, and thus the context-specific component of opportunity cost of a trial is a sunken cost paid at the outset of a trial. This inspires a second *trial-aware* hypothesis

$$\mathcal{C}_t^{\mathrm{del}} = \rho t + (\rho_\alpha - \rho)T_\alpha \,. \quad \text{(trial-aware opportunity cost)} \tag{5}$$

Equation (5) is plotted over trial time $t$ in fig. 2b(bottom). Its first term is the AR-RL contribution from the stationary opportunity cost of moment-to-moment decisions using the stationary reward rate, $\rho$ estimated with $\hat{\rho}_k^{\tau_{\mathrm{long}}}$. The second, novel term in eq. (5) is a context-specific trial cost deviation incurred at the beginning of each trial and computed as the average deviation in opportunity cost accumulated over a trial from that context ($T_\alpha$ is the average duration of a trial in context $\alpha$). This deviation fills the cost gap made by using the stationary reward rate $\rho$ in the moment-to-moment opportunity cost instead of the context-specific average reward, $\rho_\alpha$. This baseline cost derived from the orange time series in fig. 2a(bottom) vanishes in expectation, as verified through the mixed-context ensemble average reward (e.g. $\rho \equiv \sum_\alpha \rho_\alpha T_\alpha / \sum_\alpha T_\alpha$ when the context is distributed evenly among trials such that $\sum_\alpha (\rho_\alpha - \rho)T_\alpha = 0$). Thus, this opportunity cost reduces to that used in AR-RL when ignoring context, and suggests a generalization of average-adjusted value functions to account for non-stationary context. We estimate this baseline cost using $(\hat{\rho}_{k-1}^{\tau_{\mathrm{context}}} - \hat{\rho}_{k-1}^{\tau_{\mathrm{long}}})T_{k-1}$, where we have used the sample $T_{k-1}$ in lieu of the average $T_\alpha$. See fig. S1 for a signal filtering diagram that produces this estimate of eq. (5) from reward history. A main difference between the cost profiles from the two hypotheses is the cost at early times. Both the behaviour and neural recordings we analyze below seem to favor the second, trial-aware hypothesis eq. (5). We hereon employ that version in the main text, and show the results for the trial-unaware hypothesis in fig. S7.

### B. Neuroscience application: PGD in the tokens task

In this section, we apply the PGD algorithm to the "tokens task" [16]. We first give a simulated example with periodic context dynamics. We then present an application to a set of non-human primate experiments in which context variation was non-stationary [19]. For the latter, we used the decision time dynamics over trials to fit a model for each of the two subjects. We then validated the models by assessing their ability to explain (1) the concurrently recorded behaviour via their context-specific behavioural strategies and (2) the neural activity in premotor cortex (PMd) via the temporal profile of the underlying neural urgency signals.

In the tokens task, the subject must guess as to which of two peripheral reaching targets will receive the majority of tokens that randomly jump, one by one every 200ms, from a central pool initialized with a fixed number of tokens. Importantly, after the subject reports, the interval between remaining jumps contracts to once every 150ms (the "slow" condition) or once every 50ms (the "fast" condition), giving the subject the possibility to save time by taking an early guess. The interval contraction factor, $1-\alpha$, for slow ($\alpha = 1/4$) and fast ($\alpha = 3/4$) condition is parametrized $\alpha \in [0, 1]$, the incentive strength to decide early, which then serves as the task context.

In contrast to the patch leaving task example from Section A, the tokens task has many within-trial states and the state dynamics is stochastic. With the $t^{\text{th}}$ jump labelled $S_t \in \{-1, 1\}$ serving as the state, for the purposes of prediction, the history of states can be compressed into the tokens difference, $N_t = \sum_{i=1}^{t} S_i$, between the two peripheral targets with $N_0 = 0$. The dynamics of $N_t$ is an unbiased random walk (see fig. 3a), with its current value sufficient to determine the belief of a correct report, $b_t$ (computed in Methods). Since for binary rewards, $b_t$ is also the expected reward, $N_t$ is also sufficient for determining the opportunity cost of commitment, $\mathcal{C}_t^{\text{com}}$ (eq. (2)). We display this commitment cost dynamics in fig. 3b. It evolves on a lattice (gray), always starting at 0.5 (for $p = 1/2$) and ending at 0 for all $p$. We assume the agent has learned to track this commitment cost. The PGD agent uses this commitment cost, along with the estimate of the trial-aware deliberation cost, to determine when to stop deliberating and report its guess.

### 1. *A simulated example for a regularly alternating context sequence*

We first show the behaviour of the PGD algorithm in the simple case where $\alpha$ switches back and forth every 300 trials (see fig. 3). We call such segments of constant $\alpha$ 'trial blocks', with context alternating between slow ($\alpha = 1/4$) and fast ($\alpha = 3/4$) blocks. The decision space in PGD is a space of opportunity costs, equivalent to the alternative decision space formulated using beliefs [7]. In particular, one can think of the deliberation cost as the decision boundary (fig. 3b). This boundary is dynamic (see Supplemental video), depending on performance history via the estimates, $\hat{\rho}_k^{\tau_{\text{context}}}$ and $\hat{\rho}_k^{\tau_{\text{long}}}$, of the context-conditioned and stationary average reward, respectively. The result of these dynamics is effective context planning: the PGD algorithm sacrifices accuracy to achieve shorter trial duration in trials of the fast block, achieving a higher context-conditioned reward rate compared to decisions in the slow block (*c.f.* the slopes shown in the inset of fig. S2d). This behaviour can be understood by analyzing the dynamics of $\hat{\rho}_k^{\tau_{\text{context}}}$ and $\hat{\rho}_k^{\tau_{\text{long}}}$, and their effect on the dynamics of the decision time ensemble.

The two performance estimates behave differently from one another solely because of their distinct integration times. Ideally, an agent would choose $\tau_{\text{context}}$ to be large enough that it serves to average over trial-to-trial fluctuations in a context, but short enough to not average over context fluctuations. In contrast, the value of $\tau_{\text{long}}$ would be chosen large enough to average over context fluctuations. We apply those choices in this simulated example, with rounded values chosen squarely in the range in which the values inferred from the behaviour in the following application will lie. As a result of this chosen values, the context estimate $\hat{\rho}_k^{\tau_{\text{context}}}$ relaxes relatively quickly after context switches to the context-conditioned stationary average performance (dashed lines in fig. 3d), but exhibits stronger fluctuations as a result. The estimate of the stationary reward, $\hat{\rho}_k^{\tau_{\text{long}}}$, on the other hand has relatively smaller variance. This variance results from the residual zigzag relaxation over the period of the limit cycle. Given the characteristic block duration, $T_{\text{block}}$, we can more more precise. In particular, when $T_{\text{block}}$ is much less than $\tau_{\text{long}}$ ($T_{\text{block}}/\tau_{\text{long}} \ll 1$), the within-block exponential relaxation is roughly linear. Thus, the average unsigned deviation between $\hat{\rho}_k^{\tau_{\text{long}}}$ and the actual stationary reward, $\rho$, can be approximated using $1 - \exp\left[-T_{\text{block}}/\tau_{\text{long}}\right] \approx$

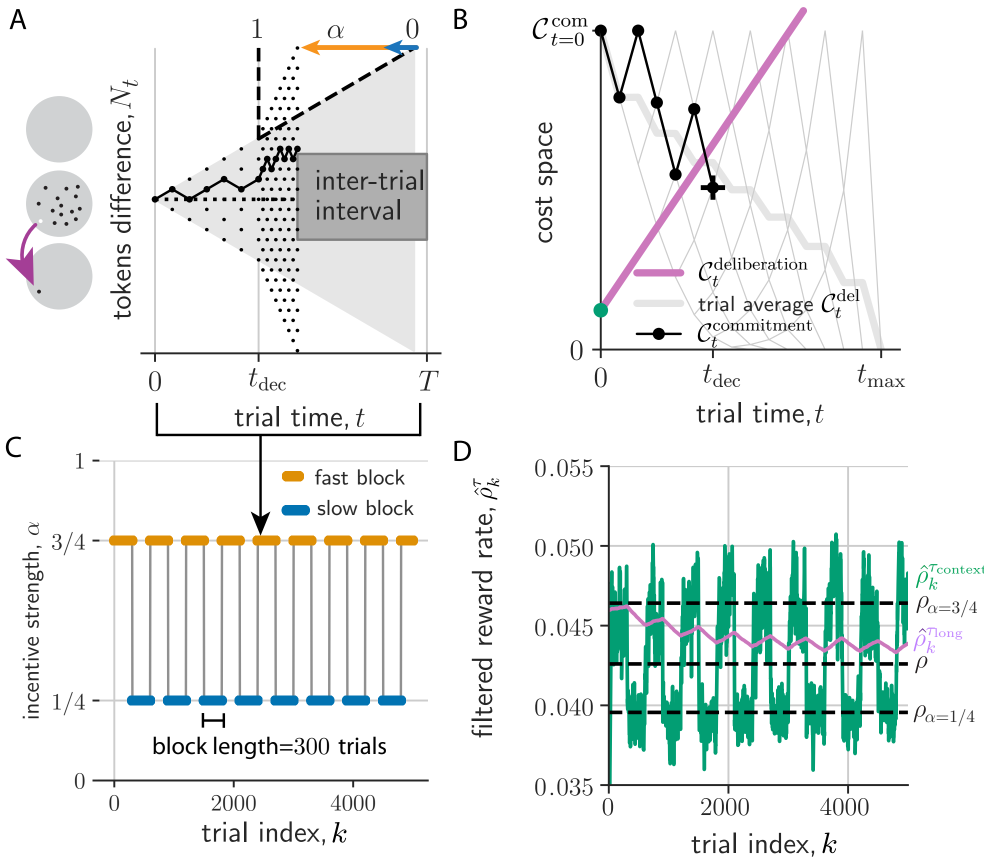



Figure 3. *PGD agent performs the tokens task for periodic context switching.* (a) A tokens task trial. Left: Tokens jump from a center to a peripheral region (gray circles). Right: The tokens difference, $N_t$, evolves as a random walk that accelerates according to $\alpha$ (here 3/4) post-decision time, $t^{\text{dec}}$. The trial duration is $T$, which includes an inter-trial interval. (b) Decision dynamics in cost space obtained from evidence dynamics in (a). Commitment cost trajectories (gray lattice; thick gray: trial-averaged) start at $\mathcal{C}_{t=0}^{\text{com}}$ and end at 0. Trajectory from (a) shown in black. $t^{\text{dec}}$ (black cross) is determined by the crossing of the commitment and deliberation cost. (c) Incentive strength switches between two values every 300 trials. (d) Expected rewards filtered on $\tau_{\text{long}}$ ($\hat{\rho}_k^{\tau_{\text{long}}}$, purple) and $\tau_{\text{context}}$ ($\hat{\rho}_k^{\tau_{\text{context}}}$, green). Black dashed lines from bottom to top are $\rho_{\alpha=1/4}$, $\rho$, and $\rho_{\alpha=3/4}$.

$T_{\text{block}}/\tau_{\text{long}} \ll 1$. This scaling fits the simulated data well (fig. S2d: inset).

The dynamics of these two performance estimates drives the dynamics of the $k$-conditioned decision time ensemble via how they together determine the deliberation cost (eq. (5); Supplemental video). For example, the mean component of this ensemble relaxes after a context switch to the context-conditioned average, while the fluctuating component remains strong due to the sequence of random walk realizations (fig. S2c). In the case of periodic context, the performance estimates and thus also the decision time ensemble relax into a noisy periodic trajectory over the period of a pair of fast and slow blocks (fig. 3d). Over this period, they exhibit some stationary bias and variance relative to their corresponding stationary averages (distributions shown in fig. S2e).

### *2. Fit to behavioural data from non-human primates and model validation*

Next, we fit a PGD agent to each of the two non-human primates' behaviour in the tokens task experiments reported in [19]. As with the above example (*c.f.* fig. 3), trials were structured in alternating blocks of two values of $\alpha$. We used the actual context-switching $\alpha$-sequence from these experiments, which, in contrast to the above example. exhibits large, irregular fluctuations in block size, primarily as a result of the experimenter adapting to fluctuations in motivation of the subject (see fig. 5a)[29].

So far, PGD has only two free parameters: the two filtering time constants, $\tau_{\text{long}}$ and $\tau_{\text{context}}$. We anticipated only a weak dependence of the fit on the $\tau_{\text{long}}$, so long as it exceeded the average duration of a handful of trial blocks enabling a sufficiently precise estimate of $\rho$. In contrast, the context filtering timescale, $\tau_{\text{context}}$, is a crucial parameter as it dictates where the PGD agent lies on a bias-variance trade-off in estimating $\rho_{\alpha_k}$, the value of which determines the context-specific contribution to the deliberation cost (eq. (2)). To facilitate the model's ability to fit individual differences, we introduce a subjective reward bias factor, $K$, that scales the rewards fed into the performance filters. We also added a tracking-cost sensitivity parameter, $\nu$, that controls $\tau_{\text{context}}$ to avoid wasting adaptation speed (see Methods for details). The latter made it possible to fit the asymmetric switching behaviour observed in the average decision time dynamics. With these four parameters, we could quantitatively match the average decision time dynamics around the two context switches (fig. 4a,b; see Methods for fitting details). A comparison of the best-fitting parameter values over the two monkeys (fig. 4c-e) suggests that the larger the reward bias (fig. 4d), the more hasty the context-conditioned performance estimate (the smaller $\hat{\tau}_{\text{context}}$), and the lower the sensitivity to the tracking cost (fig. 4e). This is consistent with the hypothesis that subjects withhold cognitive effort in contexts of higher perceived reward [8]. Inspecting the shape of the basins around the best-fitting values, we confirmed our expectation that the fitting error along the $\tau_{\text{long}}$ dimension was relatively flat above a soft lower bound (around 5000 time steps in fig. 4c; the upper bound for visually acceptable fits was imposed by the duration of the experiment). Indeed, this error basin was much wider in this dimension than along the $\tau_{\text{context}}$-dimension. Along with the correspondence in temporal statistics of the behaviour (e.g. fig. S6), the fitted model parameters for the two subjects provides a basis on which to interpret the subject differences in the results of the next section, in particular their separation on a speed-accuracy trade-off, as originating in the distinct reward sensitivity shown here.

With the models fit, we then tested them on the state-dependence of their decisions. A robust and rich representation of the behavioural statistics is the state and time-conditioned survival probability that a decision has not yet occurred. It serves as a summary of the action policy associated with a stationary strategy (see Methods for its calculation from response times). Applied equally to the decision times of both model and data, it can provide a means of comparison even in this non-stationary setting. We give this conditional probability for each of the two contexts for subject 1 in fig. 5b-e. We left the many possible noise sources underlying the behaviour out of the model in order to more clearly demonstrate the PGD algorithm. However, such noise sources would be necessary to quantitatively match the variability in the data (e.g. added noise in the performance estimates leads to larger variability in the location of the decision boundary and thus also to larger spread in these survival probability functions (not shown)). In the absence of these noise sources, we see the model underestimates the spread of probability over time and tokens state. Nevertheless,

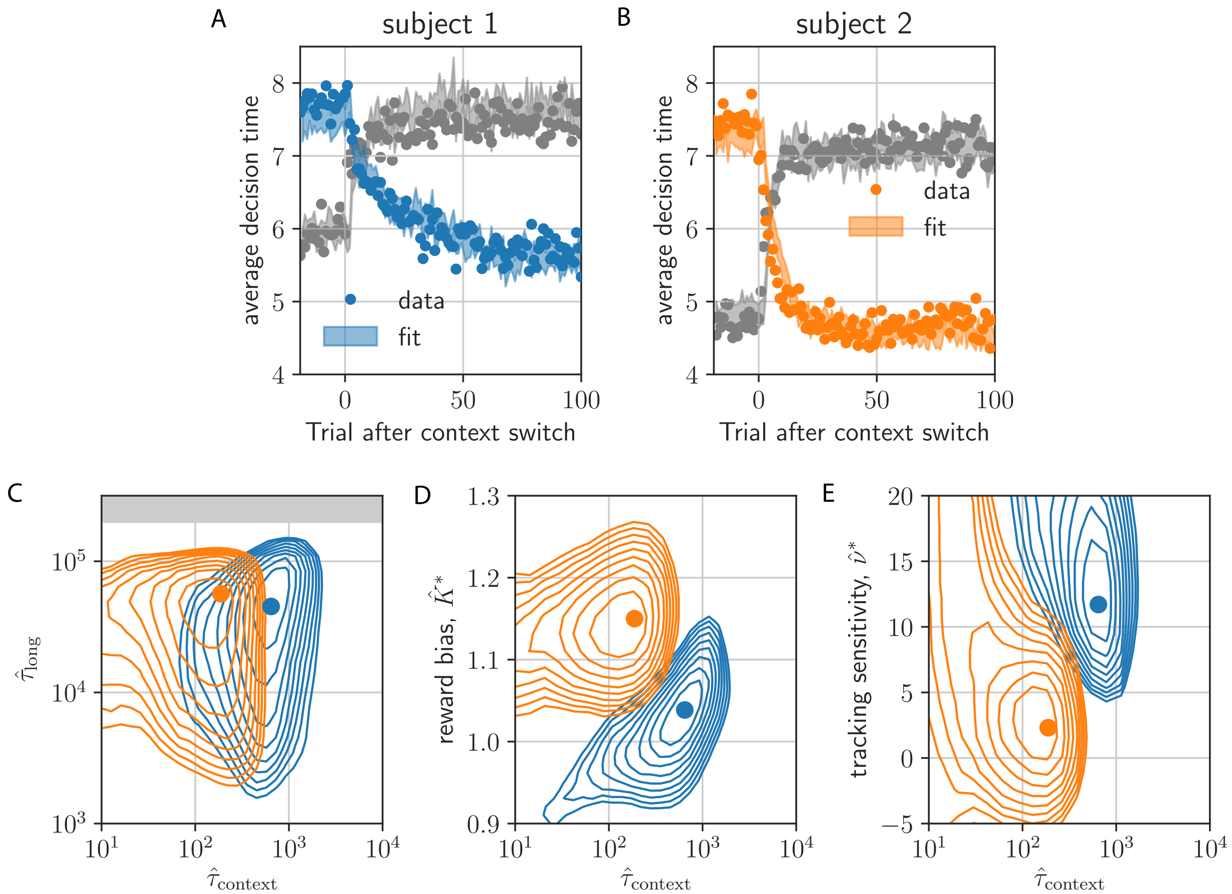



Figure 4. *Model fit.* (a,b) decision times (dots) aligned on the context-switching event type (fast-to-slow in gray; slow-to-fast in color) and averaged. Shaded regions are the standard error bounds of the models' average decision times. (c) Error evaluated on a $(\hat{\tau}_{\text{context}}, \hat{\tau}_{\text{long}})$-plane cut through the parameter space at the best-fitting $\nu = \hat{\nu}^*$ and $K = \hat{K}^*$ (gray area indicates timescales within an order of magnitude of the end of the experiment). Contours show the first 10 contours incrementing by 0.01 error from the minimum (shown as a circle marker). Colors refer to subject, as in (a) and (b). (d) Same for $(\hat{\tau}_{\text{context}}, \hat{K})$ at $\hat{\tau}_{\text{long}} = \hat{\tau}^*_{\text{long}}$ and $\nu = \hat{\nu}^*$. (e) Same for $(\hat{\tau}_{\text{context}}, \hat{\nu})$ at $\hat{\tau}_{\text{long}} = \hat{\tau}^*_{\text{long}}$ and $K = \hat{K}^*$.

the remarkably smooth average strategy is well captured by the model (white dashed lines in fig. 5c-e). Specifically, policies approximately decide once either of the peripheral targets receive a certain number of tokens. Comparing results across context, we find that fast block strategies (fig. 5d,e) exhibit earlier decision times relative to slow block strategies (fig. 5d,e) in both model and data. The strategies for subject 2 are qualitatively similar, but shifted to earlier times relative to subject 1 (fig. S3). Our model explains this subject difference as resulting from subject 2's larger reward bias and faster context integration (*c.f.* fig. 4d). The correspondence between model and data over the many token states in fig. 5b-e is remarkable given that the model has essentially only a single, crucial degree of freedom ($\tau_{\text{context}}$), *a priori* unrelated to how decision times depend on token state.

To better understand where both the data and the PGD agent lie in the space of strategies for the tokens task, we computed reward-rate (AR-RL) optimal solutions for a given fixed

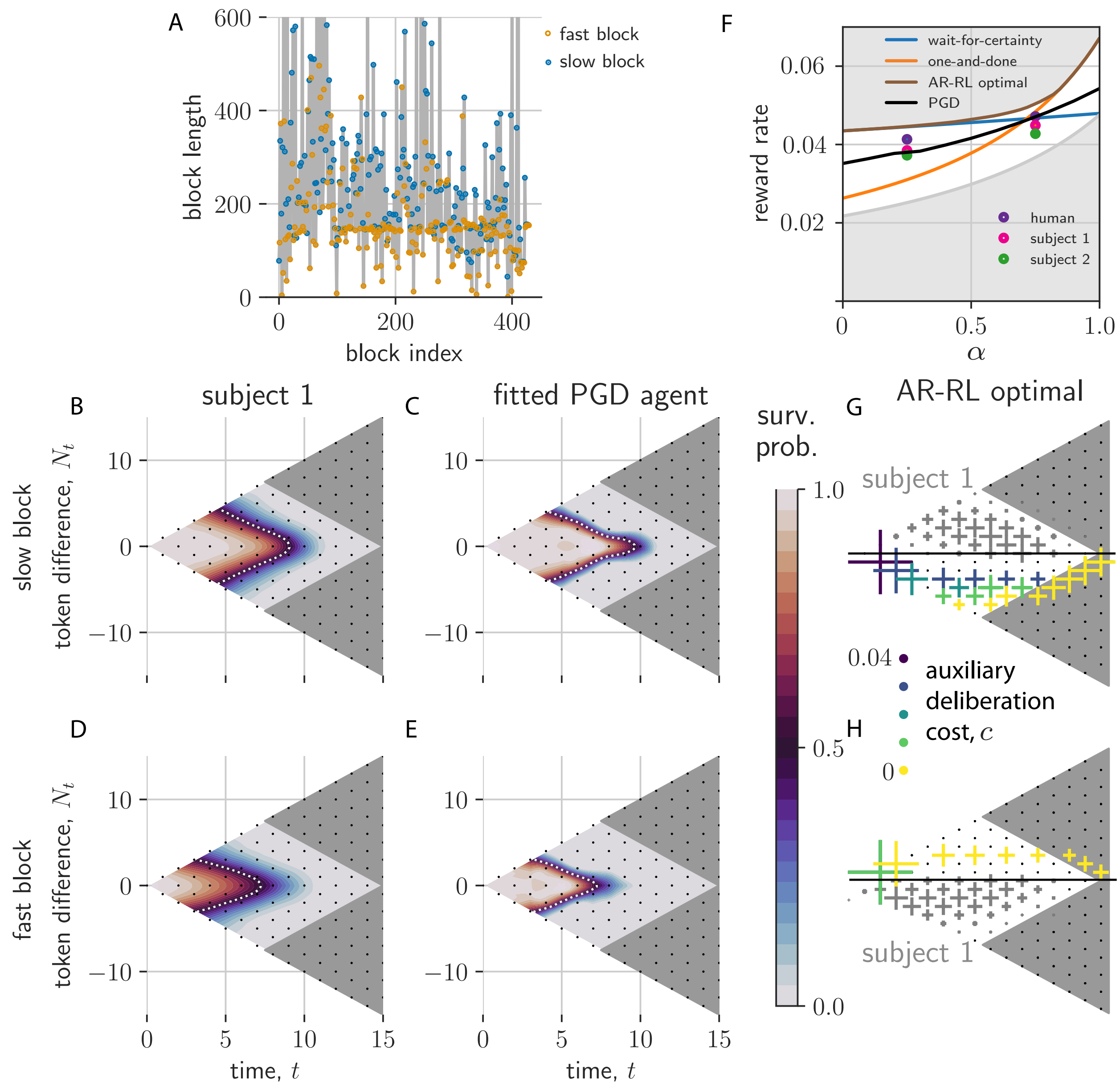



Figure 5. *Comparison of PGD to NHP data for non-stationary $\alpha$-dynamics from Ref. [19].* (a) Block length sequence used in the experiment (*c.f.* fig. 3c). (b-e) Interpolated state-conditioned survival probabilities, $P(t^{\text{dec}} = t|N_t, t)$, over slow (b,c) and fast (d,e) blocks. White dotted lines show the $P(t^{\text{dec}} = t|N_t, t) = 0.5$ contour. (f) Shown is the reward rate as a function of incentive strength, $\alpha$ (wait-for-certainty strategy shown in blue; one-and-done strategy shown in orange). We additionally show the slow and fast context-conditioned reward rates for the two primates as well as a reference expert human, and the PGD solution (black line). Reward rates for the human and non-human primates are squarely in between the best (brown) and uniformily random (gray) strategy. (g,h) State-conditioned decision time frequencies (cross size) from AR-RL optimal decision boundaries across different values of the auxiliary deliberation cost (colored crosses) for slow (g) and fast (h) conditions. Only samples with $N_t < 0$ and $N_t > 0$, respectively, are shown. The reflected axes shows as gray crosses the subject's state-conditioned decision time frequencies for comparison.

context, $\alpha$ (here $\alpha \in [0,1]$), using the same approach as [7]. Iterating Bellman's equation for this decision problem backwards from the end of the trial provides the optimal value functions from which the optimal policy and its reward rate can be obtained (see Methods for details). The optimal reward rate as a function $\alpha$ is shown in fig. 5f. The optimal strategies generating these reward rates interpolate from the wait-for-certainty strategy at low $\alpha$ to the one-and-done strategy [30] at high $\alpha$. The $\alpha$-conditioned reward rates achieved by the two primates and a reference human [31] are also shown in fig. 5f. They fall conspicuously below the optimal strategy, and, as expected, above the strategy that picks one of the three actions (report left, report right, and wait) at random. Given the good match in behaviour between the PGD model and data (*c.f.* fig. 5b,e), this intermediate performance is shared by PGD's solution (black line).

What are the differences in decision times that underlie these performance differences? Both the fitted PGD model and the primate behaviour resolve residual ambiguity ($N_t \approx 0$) at intermediate trial times (fig. 5b-e). In contrast, the optimal strategies give no intermediate decision times at ambiguous ($N_t \approx 0$) states, invariably waiting until the ambiguity resolves (see fig. 5g,h). To assess the extent of this difference, and for comparison with previous work [7], we added to the reward objective a constant auxiliary deliberation cost rate, $c$, incurred up to the decision time in each trial. We find that the resulting optimal strategies lack intermediate decision times at ambiguous states for all $c > 0$ and in fact over the entire $(\alpha, c)$-plane ( see fig. S9 for the complete dependence). This holds also under the addition of a movement cost, i.e. a constant cost incurred by either of the reporting actions (data not shown). Thus, whereas optimal policies shift around the edges of the relevant decision space as $\alpha$ or $c$ is varied, the PGD policy lies squarely in the bulk, tightly overlaying the policy extracted from the data. We conclude that the context-conditioned strategies of the non-human primates in this task are well-captured by PGD, while having little resemblance to the behaviour that would maximize reward rate with or without a fixed auxiliary deliberation cost rate.

### *3. Neural urgency and context-dependent opportunity cost*

So far, we have fit and analyzed the PGD model with respect to recorded behaviour. Here, we take a step in the important direction of confronting the above theory of behaviour with the neural dynamics that we propose drive it. The proposal for the tokens task mentioned at the end of the introduction has evidence strength and urgency combining in PMd, whose neural dynamics implements the decision process. In fig. 6a, we restate in a schematic diagram an implementation of this dynamics that includes a collapsing decision boundary. In the one-dimensional belief space for the choice (fig. 6a(top)) [7, 32], the rising belief collides with the collapsing boundary to determine the decision time. In the equivalent commitment and deliberation cost formulation developed here (fig. 6a(middle)), the falling commitment cost collides with the rising deliberation cost. The collapsing boundary in belief space can be parametrized as $C - u_t$, where $C$ is the initial strength of belief, e.g. some desired confidence, that is lowered by a growing function of trial time $u_t > 0$. The decision criterion is then $b_t > C - u_t$, where $b_t$ is the belief, i.e. the probability of a correct report. For AR-RL optimal policies, $u_t$ emerges from value maximization and thus has a complicated dependence on the opportunity cost sequence, $\mathcal{C}_t^{\mathrm{del}}$. For PGD, in contrast, $C$ is interpreted as the maximum reward $r_{\max}$ and $u_t$ is identically $\mathcal{C}_t^{\mathrm{del}}$. For a linear neural encoding model in which belief, rather than evidence, is encoded in neural activity, the sum of the encoded belief $\tilde{b}_t$ and the encoded collapsing boundary, $\tilde{u}_t$, evolve on a one-dimensional choice manifold. According to the proposal, when this sum becomes sufficiently large (e.g. $\tilde{b}_t + \tilde{u}_t > \tilde{C}$ for some threshold $\tilde{C}$), PMd begins to drive the activity in downstream motor areas towards the associated response.

Neural urgency was computed from the PMd recordings of [19] in [33]. This computation relies on the assumption that while a single neuron's contribution to $\tilde{b}_t$ will depend on its selectivity for choice (left or right report), the urgency $\tilde{u}_t$ is a signal arising from a population-level drive to all PMd neurons, irrespective of their selectivity. Thus, $\tilde{u}_t$ can be extracted from neural recordings by conditioning on zero-evidence states ($\tilde{b}_t = 0$) and averaging over cells. In [33], error bars were computed at odd times via bootstrapping; data at even times was obtained by interpolating between $N_t = \pm 1$; and data was pooled from both subjects. We have excluded times at which firing rate error bars exceed the range containing predictions from both blocks. To assess the correspondence of the components of the deliberation cost developed here and neural urgency, in fig. 6b we replot their result (*c.f.* fig.8b of [33]). We overlay the mean (+/- standard deviation) of the opportunity cost sequence, $\mathcal{C}_t^{\mathrm{del}}$ (shaded area in fig. 4; averaged over all trials produced by applying the two fitted PGD models on the data sequence and conditioning the resulting average within-trial deliberation cost on context). To facilitate our qualitative comparison, we convert cost to spikes/step simply by adjusting the y-axis of the deliberation cost. The observed urgency signals then lie within the uncertainty of the context-conditioned deliberation cost signals computed from the fitted PGD models. There are multiple features of the qualitative correspondence exhibited in fig. 6b: (1) the linear rise in time; (2) the same slope across both fast and slow conditions; and (3) the baseline offset between conditions, where the fast condition is offset to higher values than the slow condition. Such features would remain descriptive in the absence of a theory. With the theory we have presented here, however, each has their respective explanations via the interpretation of urgency as the opportunity cost of deliberation: (1) the subject uses a constant cost per token jump, (2) this cost rate refers to moment-to-moment decisions, irrespective of context, that is reflective of the use of the context-agnostic stationary reward, and (3) trial-aware planning over contexts leads to an opportunity cost baseline offset with a sign given by the reward rate deviation $\rho_\alpha - \rho$ with respect to the stationary average, $\rho$.

Up to now, the computational and neural basis for urgency has remained largely unexplored in normative approaches, which also typically say little about adaptation effects (see [34] for a notable exception). In summary, we exploited the adaptation across context switches to learn the model and explained earlier responses in high reward rate contexts as the result of a higher opportunity cost of deliberation. While this qualitative effect is expected, we go beyond existing work by quantitatively predicting the average dependence on both time and state (fig. 5b-e) as well as the qualitative form of urgency signal (fig. 6b). Taken together, the data is thus consistent with our interpretation that neural activity underlying context-conditioned decisions is gated by opportunity costs reflective of a trial-aware timescale hierarchy computed using performance estimation on multiple timescales.

## DISCUSSION

We have proposed PGD, a heuristic decision-making algorithm for continuing tasks that gates deliberation based on performance. We constructed a foraging example for which

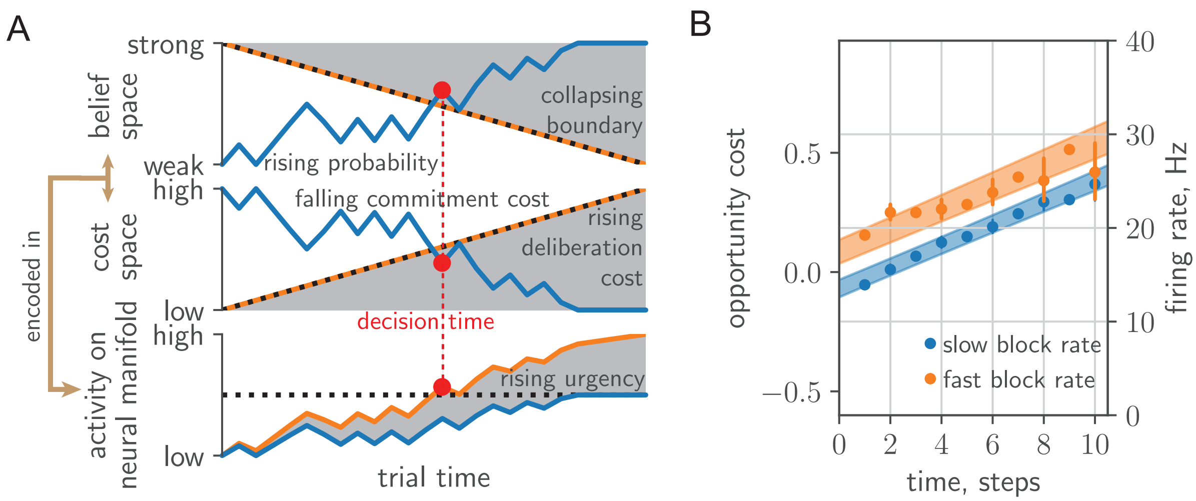



Figure 6. *Comparing neural urgency and collapsing decision boundaries.* (a) Top: Rising belief (blue) meets collapsing decision boundary (black dashed) in belief space. Middle: Falling commitment cost (blue) meets rising deliberation cost (black-dashed) in cost space. Bottom: Belief/commitment cost is encoded (blue) into a low-dimensional neural manifold, with the addition of an urgency signal (orange) (*c.f.* fig.8 in [7]). The decision (red circle) is taken when the sum passes a fixed threshold (black-dashed). (b) Deliberation cost maps onto the urgency signal extracted from zero-evidence conditioned cell-averaged firing rate in PMd.

PGD is the optimal strategy with respect to the average-adjusted value function of average-reward reinforcement learning (AR-RL). While this will not be true in general, PGD does strike a balance between strategy complexity and return. The PGD decision rule does not depend on task specifics and exploits the stationarity of the environment statistics while simultaneously hedging against longer term non-stationarity in reward context. It does so by splitting the problem into two separate components—learning the statistics of the environment in order to form the opportunity cost of commitment, and tracking one's own performance in that environment in order to form the opportunity cost of deliberation. Within our current understanding of how the cortico-basal ganglia system supports higher-level decision-making [35], this latter cost is proposed as arising from performance estimated on multiple, behaviourally-relevant timescales that are broadcast to multiple, lower-level decision-making areas to gate the speed of their respective attractor-based decision-making dynamics (models of the latter are well-studied [32, 36, 37]). Consistent with this picture, PGD's explanatory power was borne out at both the behavioural and neural levels for the tokens task data we analyzed. In particular, a deliberation cost constructed from trial-aware planning was supported independently by both modalities. We used behavioural data to fit and validate the theory, and neural recordings to provide evidence of one of the neural correlates it proposes: the temporal profile of neural urgency.

*Scientific and clinical implications* In our proposal, we have linked two important and related, but often disconnected fields: the systems neuroscience of the neural dynamics of decision-making and the cognitive neuroscience of opportunity cost and reward sensitivity. The view that tonic dopamine encodes average reward is two decades old [3]. However, the existence of a reward representation decomposed by timescale has received increasing empirical support in recent years, from cognitive results [38–40] to a recent unified view of how dopamine encodes reward prediction errors using multiple discount factors [41, 42] and of dopamine as encoding both value and uncertainty [43]. Dopamine's effect on time perception has been proposed [44] and has empirical support [45], but the mechanism by which its putative effect on decision speed is implicated in the neural dynamics of the decision-making areas driving motor responses was unknown. Our theory fills this explanatory gap by considering dynamic evidence tasks and parametrizing urgency using a multiple-timescale representation of performance. One candidate for the latter's neural implementation is in the complex spatio-temporal filtering of dopamine via release-driven tissue diffusion and integration via DR1 and DR2 binding kinetics [46]. Subsequent neural filtering and computation by striatal network activity could also play a role [47]. The study of spatiotemporal filtering of dopamine is increasingly accessible experimentally [48, 49] and provides an exciting direction for multi-scale analysis of behaviour. Our proposal that urgency is the means by which the neural representation of reward ultimately affects neural dynamics in decision-making areas frames a timely research question on which these questions could shed light.

While we applied PGD to decisions playing out in PMd, a decision-making area relevant to arm movements, PGD may also be relevant to other kinds of decisions. For instance, a large body of work has studied decisions playing out in lateral intraparietal cortex using random dot motions tasks. One seminal study identified an urgency signal with the same properties: a linear dependence at early trial times and an offset with sign given by the reward rate deviation across two and four-choice trials, which serve as the two contexts [17]. In contrast to the tokens task, however, context was sampled randomly and thus its dynamics lacked temporal correlation. In this case, we might expect that a pair of performance filters, one for each context, to track the reward history in two parallel streams, each updated only when in their respective context. In this case, our theory would predict that the ratio of slopes of urgency reflect the ratio of context-conditioned reward rates. An estimation given in the Methods for this data [17] agrees to within 20% error, providing preliminary support for our theory. Testing the generality of our theory using tailored experiments in this other setting is an important next step.

Urgency may play a role in both decision and action processes, potentially providing a transdiagnostic indicator of a wide range of cognitive and motor impairments in Parkinson's disease and depression [50]. Our theory offers a means to ground these diverse results in neural dynamics by formulating opportunity cost estimation as the underlying causal factor linking vigor impairments (e.g. in Parkinson's disease) and dysregulated dopamine signalling in the reward system [50–52]. We provide a concrete proposal for a signal filtering system that extracts a context-sensitive opportunity cost from a reward prediction error sequence putatively encoded by dopamine. Neural recordings of basal ganglia provide a means to identify the neural substrate for this system.

*Commitment cost estimation* Beyond the estimation of the opportunity cost of deliberation, we assumed that the agent had a precise estimate of the expected reward, which it used to compute the within-trial commitment cost. For the tokens task, a recorded signal in dorsal lateral prefrontal cortex of non-human primates correlates strongly with belief [20], equivalent to the expected reward for binary rewards). How this quantity is computed by neural systems is not currently known. However, for a general class of tasks, a generic, neurally plausible means to learn the expected reward is via distributional value codes [43]. For example, the Laplace code is a distributional value representation that uses an ensemble of units over a range of temporal discount factors and reward sensitivities [53]. The authors show that expected reward is linearly decodeable from this representation.

*Experimental predictions* A feature of our decision-making theory is that it is highly vulnerable to falsification. First, with regards to behaviour via the shape of the action policy using our survival probability representation (*c.f.* fig. 5b-e,g,h), PGD varies markedly with reward structure and thus provides a wealth of predictions for how observed behaviour should be altered by it. For example, a salient feature of the standard tokens task is its reflection symmetry in the tokens difference, $N_t$. We can break this symmetry for which the theory predicts a distinctly asymmetric shape (fig. S10; for details see Methods). Our theory is also prescriptive for neural activity via the temporal profile of neural urgency. The slope of $\mathcal{C}_t^{\mathrm{del}}$ remained fixed across blocks for relatively short block lengths used in the data analyzed here. In the opposite limit, $T_{\mathrm{block}}/\tau_{\mathrm{long}} \gg 1$, $\rho_k^{\tau_{\mathrm{long}}}$ approaches $\rho_\alpha$ except when undergoing large, transient excursions after context switches. Thus, the deliberation cost is given by the first component in eq. (5) most of the time, with the context specific reward rate as the slope. One simple prediction is that the slope of urgency should exhibit increasing variation as the duration of the blocks increases.

*Reinforcement learning theory* Our work impacts reinforcement learning theory by suggesting how to generalize average-adjusted value functions to time-varying opportunity cost of time in a way that reduces to AR-RL when context is not tracked. This further develops episodic AR-RL in the continuing task setting, which has received relatively little attention from AR-RL machine learning research, and yet is central to experimental neuroscience. The epistemic perspective entailed in the estimation of these costs parallels a recent epistemic interpretation of the discount-reward formulation as encoding knowledge about the volatility of the environment [54].

Our work also suggests a new class of reinforcement learning algorithms between model-based and model-free: only parts of the algorithm need adjustment upon task structure variation. This is reminiscent of how the effects of complex state dynamics are decoupled from reward when using a successor representation [55], but tailored for the average-reward rather than the discount-reward formulation. We have left a detailed algorithmic analysis of PGD to future work, but expect performance improvements, as with successor representations, in settings where decoupling the learning of environment statistics from the learning of reward structure is beneficial.

*Comparison with humans* In the space of strategies, PGD lies in a regime between fully exploiting assumed task knowledge (average-case optimal) and assumption-free adaptation (worst-case optimal). Highly incentivized human behaviour is likely to be more structured than PGD because of access to more sophisticated learning. While some humans land on the optimal one-and-done policy in the fast condition when playing the tokens task [56], most do not. The human brain likely has all the components needed to implement PGD. Nevertheless, the situations in which we actually exploit PGD, if any, are as yet unclear. In particular, how PGD and AR-RL relate to existing behavioural models tailored to explain relative-value, context-dependent decision-making in humans [4], such as scale and shift adaptation[57], is an open question. Whether or not PGD is built into our decision-making, the question remains if PGD is optimal with respect to some bounded rational objective. In spite of the many issues with the latter approach [58], using it to further understand the computational advantages of PGD is an interesting direction for future work.

Despite our putative access to sophisticated computation, humans still exhibit measurable bias in how we incorporate past experience [59]. One simple example is the win-stay/lose-shift strategy, a more rudimentary kind of performance-gated decision-making than PGD, which explains how humans approach the rock-paper-scissors game [60]. In that work, numerical experiments demonstrated that this strategy outperforms at a population level the optimal Nash equilibrium for this game, demonstrating that the use of such seemingly sub-optimal strategies can confer a surprising evolutionary advantage. This example supports the claim that relatively simple and nimble strategies such as PGD make for attractive candidates when acknowledging that a combination of knowledge and resource limitations over task, development, and evolutionary timescales have shaped decision-making in non-stationary environments.

## METHODS

Code for simulations and main figure generation (written in Python 3) is publicly accessible as a online repository: https://github.com/mptouzel/dyn_opp_cost/.

### Patch leaving task

We devised a mathematically tractable patch leaving task for which PGD learning is optimal with respect to the average-adjusted value function. Here the value is simply the return from the patch. This value function is related, but not equivalent to the marginal value of optimal foraging, for which the decision rule is $\mathcal{C}_t^{\mathrm{del}} > r_{\max} - \mathcal{C}_t^{\mathrm{com}} = \bar{r}_t$ [5]). This choice of task allowed us to compare PGD's convergence properties relative to conventional AR-RL algorithms that make use of value functions. In contrast to PGD, the latter requires exploration. For a comparison generous to the AR-RL algorithm, we allowed it to circumvent exploration by estimating the value function from off-policy decisions obtained from the PGD algorithm using the same learning rate. We then compared them to PGD using their on-policy, patched-averaged reward. This made for a comparison based solely between the parameters of the respective models. If we did not allow for this, the ar-RL algorithms would have to find good learning signals by exploring. In any form, this exploration would lead them converge substantially slower. This setting thus provides a lower bound on the convergence times of the AR-RL algorithm.

In this task, the subject randomly samples (with replacement) $d$ patches, each of a distinct, fixed, and renewable richness defined by the maximum return conferred. These maximum returns are sampled before the task from a richness distribution, $p(r_{\max})$, with $r_{\max} > 0$ and are fixed throughout the experiment. The trials of the task are temporally extended periods during which the subject consumes the current patch. After a time $t$ in a patch, the return is defined $r(t) = r_{\max}(1 - (\lambda t)^{-1})$. This patch return profile, $1 - (\lambda t)^{-1}$, is shared across all patches and saturates in time with rate $\lambda$, a parameter of the environment that sets the reference timescale. The return diverges negatively for vanishing patch leaving times for mathematical convenience, but also evokes situations where leaving a patch soon after arriving is prohibitively costly (e.g. when transit times are long). A stationary policy is then a leaving time, $t_s$, for each of $d$ patches, where the $s$-subscript indexes the patch. Given any policy, the stationary reward rate for uniformly random sampling of patches is then defined as

$$\rho = \sum_{s=1}^{d} r_s(t_s) \Big/ \sum_{s=1}^{d} t_s \ . \tag{6}$$

We designed this task to (1) emphasize the speed-return trade-off typical in many deliberation tasks, and (2) have a tractable solution with which to compare convergence properties of PGD and AR-RL value function learning algorithms.

A natural optimal policy is the one that maximizes the average-adjusted trial return, $Q(r,t) = r - \rho t$. Given the return profile we have chosen, the corresponding optimal decision time, $t_s^*$, in the $s$th patch obtained by maximizing $r - \rho t$ is $t_s^* = \sqrt{r_{\max,s}/(\lambda\rho)}$, which scales inversely with the reward rate so that decision times are earlier for larger reward rates, because consumption (or more generally deliberation) at larger reward rates costs more. We chose this return profile such that stationary PGD learning gives exactly the same decision times: the condition $\mathcal{C}_t^{\mathrm{del}} = \mathcal{C}_t^{\mathrm{com}}$ for patch $s$ here takes the form $\rho t_s = r_{\max,s}/(\lambda t_s)$. Thus, they share the same optimal reward rate, $\rho^*$. Using $t_s^*$ for each patch in eq. (6) gives a self-consistency equation for $\rho$ with solution $\rho^* = \lambda\mu_1^2/4\mu_{1/2}^2$, where $\mu_n = \langle r_{\max}^n\rangle_{p(r_{\max})}$ (we have assumed $d$ is large here to remove dependence on $s$). Described so far in continuous time, the value function was implemented in discrete time such that the action space is a finite set of decision times selected using the greedy policy, $t^* = \operatorname{argmax}_t \hat{Q}(r,t)$, where $\hat{Q}(r,t)$ is the estimated trial return. As a result, there is a finite lower bound on the performance gap, i.e. the relative error, $\epsilon = (\rho^* - \rho)/\rho^* > 0$ for the AR-RL algorithm. Approaching this bound, convergence time for both PGD and AR-RL learning is limited by the integration time $\tau$ of the estimate $\hat{\rho}_k^\tau$ (*c.f.* eq. (8)) of $\rho$. We note that PGD learns faster in all parameter combinations tested. To demonstrate the insensitivity of PGD to the state space representation, at $5 \times 10^5$ time steps into the experiment we shuffled the labels of the states. PGD is unaffected, while the value function-based AR-RL algorithm is forced to relearn and in fact does so slower than in the initial learning phase, due to the much larger distance between two random samples, than between the initial values (chosen near the mean) and the target sample.

## Filtering performance history

For unit steps of discrete time, the step-wise update of the performance estimate, $\hat{\rho}_t^\tau$, is

$$\hat{\rho}_t^\tau = (1-\beta)\hat{\rho}_{t-1}^\tau + \beta R_t \;, \tag{7}$$

with $\beta = 1/(1+\tau)$ called the learning rate, and $\tau$ the characteristic width of the exponential window of the corresponding continuous time filter over which the history is averaged. We add $\tau$ as a superscript when denoting the estimate to indicate this. Exceptionally, here $t$ indexes absolute time rather than trial time. Note that a continuous-time formulation of the update is possible via an event-based map given the decision times in which the reward event sequence is given as a sum of delta functions. In either case, to leading order in $\beta$, $\hat{\rho}_t^\tau \approx \beta \sum_i^t R_i$, i.e. the filter sums past rewards. Thus, when $\tau \sim \mathcal{O}(t) \gg 1$, $\beta \sim \mathcal{O}(1/t) \ll 1$ and so $\hat{\rho}_t^\tau \approx \beta \sum_i^t R_i \to \rho$ when $t$ is large.

The rewards in this task are sparse: $R_t = 0$ except when a trial ends and the binary trial reward $R_k$ (1 or 0) is received. A cumulative update of eq. (7) that smooths the reward uniformly over the trial duration and is applied once at the end of each trial is thus more compuatationally efficient. Resolving a geometric series leads to the cumulative update [8, 28]

$$\hat{\rho}_k^\tau = (1-\beta)^{T_k}\hat{\rho}_{k-1}^\tau + (1-(1-\beta)^{T_k})\rho_k^{\mathrm{trial}} \;, \tag{8}$$

where the smoothed reward, $\rho_k^{\mathrm{trial}} = R_k/T_k$, can be interpreted as a trial-specific reward rate. The initial estimate, $\hat{\rho}_0^\tau$, is set to 0. Exceptionally, $\hat{\rho}_1^\tau = R_1/T_1$, after which eq. (8) is used.

Using the first finite sample as the first finite estimate is both more natural and robust than having to adapt from zero. We will reuse this filter for different $\tau$ and denote the filtered estimate from its application with a $\tau$-superscript, $\hat{\rho}_k^{\tau}$. For example, the precision of $\hat{\rho}_k^{\tau_{\text{long}}}$ as an estimate of a stationary reward rate $\rho$ is set by how many samples it averages over, which is determined by the effective length of its memory given by $\tau_{\text{long}}$. Since we assume the subject has learned the expected reward, $\bar{r}_t$, we use it instead of $R_k$ when computing $\rho_k^{\text{trial}}$.

### Tokens task: a random walk formulation

The tokens task is a continuing task of episodes (here trials), which can be formulated using the token difference, $N_t$. Each trial effectively presents to the agent a realization of a finite-length, unbiased random walk, $\boldsymbol{N}_{t_{\max}} = (N_0, \ldots, N_{t_{\max}})$ with $N_t = \{-t, \ldots, t\}$ and $N_0 = 0$. We express time in units of these steps. The agent observes the walk and reports its prediction of the sign of the final state, $\text{sign}(N_{t_{\max}}) = \pm 1$ ($t_{\max}$ is odd to exclude the case it has no sign). The time at which the agent reports is called the decision time, $t^{\text{dec}} \in \{0, 1, \ldots, t_{\max}\}$. For a greedy policy, $\text{sign}(N_t)$ can be used as the prediction (and the reporting action selected randomly if $N_{t^{\text{dec}}} = 0$). The decision-making task then only involves choosing when to decide. In this case, the subject receives reward $R = \Theta(N_{t_{\max}} N_{t^{\text{dec}}})$ at the end of the random walk, i.e. a unit reward for a correct prediction, otherwise nothing ($\Theta$ is the Heaviside function: $\Theta(x) = 1$ if $x > 0$, zero otherwise).

An explicit action space beyond decision time is not necessary for the case of greedy actions. It can nevertheless be specified for illustration in an Markov decision process (MDP) formulation: the agent waits ($a_t = 0$ for $t < t^{\text{dec}}$) until it reports its prediction, $a_{t^{\text{dec}}} = \pm$, after which actions are disabled and the prediction is stored in an auxiliary state variable used to determine the reward at the end of the trial. A MDP formulation for a general class of perceptual decision-making tasks, including the tokens and random dots task, is given in Methods).

Perfect accuracy in this task is possible if the agent reports at $t_{\max}$ since $R = \Theta(N_{t_{\max}}^2) = 1$. The task was designed to study reward rate maximizing policies. In particular, the task has additional structure that allows for controlling what this optimal policy is through the incentive to decide early, $\alpha$, incorporated into the trial duration for deciding at time $t$ in the trial,

$$T(t) = t + (1 - \alpha)(t_{\max} - t) + T_{\text{ITI}}. \tag{9}$$

Here, a dead time between episodes is added via the inter-trial interval, $T_{\text{ITI}}$, to make suboptimal the strategy of predicting randomly at the trial's beginning. We emphasize that it is through the trial duration that $\alpha$ serves as a task parameter controlling the strength of the incentive to decide early. When $\alpha$ is fixed, we denote the corresponding optimal stationary reward rate, $\rho_\alpha$, obtained from the reward rate maximizing policy. This policy shifts from deciding late to deciding early as $\alpha$ is varied from 0 to 1 (*c.f.* fig. S9f,g).

We consider a version of the task where $\alpha$ is variable across two episode types, a slow ($\alpha = 1/4$) and fast ($\alpha = 3/4$) type. The agent is aware that the across-trial $\alpha$ dynamics are responsive (maybe even adversarial), whereas the within-trial random walk dynamics (controlled by the positive jump probability, here $p = 1/2$) can be assumed fixed (see the next section for how $p$ factors into the expression for the expected reward, $\bar{r}_t$.

### Expected trial reward for the tokens task

We derived and used an exact expression for the expected reward in a trial of the tokens task. We derive that expression here as well as a simple approximation. The state sequence is formulated as a $t_{\max}$-length sequence of random binary variables, $\boldsymbol{S}_{t_{\max}} = (S_1, \dots, S_{t_{\max}})$, $S_t = \pm 1$, $i = 1, 2, \dots, t_{\max}$. Consider a simple case in which each is an independent and identically distributed Bernoulli sample, $P(s) = p^{\frac{1+s}{2}}(1-p)^{\frac{1-s}{2}}$, for jump probability $p \geq 1/2$. The distribution of $\boldsymbol{S}_{t_{\max}}$ is then

$$P(\boldsymbol{s}_{t_{\max}}) = \prod_{i=1}^{t_{\max}} P(s_i) \; . \tag{10}$$

We will use this distribution to compute expectations of quantities over this space of trajectories, namely the sign of $N_t = \sum_{i=1}^{t} S_i$, for some $0 \leq t \leq t_{\max}$ and in particular the sign of the final state, $\xi := \mathrm{sgn}(N_{t_{\max}}) \in \{+, -\}$ given $N_t = n$. Note that $N_t$ is even if $t$ is even and same with odd values. We remove the case of no sign in $N_{t_{\max}}$ by choosing $t_{\max}$ to be odd.

First, consider predicting $\mathrm{sgn}(N_t)$ with no prior information. The token difference, $-t \leq N_t \leq t$, appears directly in $P(\boldsymbol{s}_{t_{\max}})$. Marginalizing (here just integrating out) the additional degrees of freedom leads to a binomial distribution in the number of $S_i$ for $i \leq t$ for which $S_i = +1$, $N_t^+ = \sum_{i=1}^{t} \Theta(s_i) = (t + N_t)/2$,

$$P(N_t^+ = n) = \binom{t}{n} p^n (1-p)^{t-n} \; , \tag{11}$$

with $n \in \{0, \dots, t\}$ and $N_t = 2N_t^+ - t$. Thus, the probability that $N_t > 0$, i.e. $N_t^+ > t/2$, is

$$P(N_t > 0) = \sum_{n=0}^{t} \binom{t}{n} p^n (1-p)^{t-n} \Theta(n - t/2) \; . \tag{12}$$

Now consider predicting $\xi = \mathrm{sgn}(N_{t_{\max}})$, given the observation $N_t = n$. Define $t' = t_{\max} - t$ as the remaining time steps to the predicted time and $N_{t'} = \sum_{i=t+1}^{t_{\max}} s_i$, i.e. the total count in the remaining part of the realization. Then the probability of $\xi = +$ conditioned on the state $N_t = n$, denoted $p_{n,t}$, is defined in the same way as $P(N_t > 0)$,

$$p_{n,t}^+ := P(\xi = + | N_t = n) = \sum_{n'=0}^{t'} \binom{t'}{n'} p^{n'} (1-p)^{t'-n'} \Theta(n' - (t' - n)/2) \; . \tag{13}$$

where $N_{t'}^+ = n'$ is the number of positive jumps in the remaining $t' = t_{\max} - t$ steps and we have used $N_{t_{\max}} = N_t + N_{t'} = N_{t'}^+ - (t' - N_t)/2$. The $\Theta(n' - (t' - n)/2)$ factor effectively changes the lower bound of the sum to $\max\{0, \lceil (t' - n)/2 \rceil\}$, where $\lceil \cdot \rceil$ rounds up. If $\lceil (t' - n)/2 \rceil \leq 0$ then $p_{n,t}^+ = 1$ since the sum is over the domain of the distribution, which is normalized. Otherwise, the lower bound is $\lceil (t' - n)/2 \rceil$, and the probability of $\xi = +1$ is

$$p_{n,t}^+ = \sum_{n'=\lceil (t'-n)/2 \rceil}^{t'} \binom{t'}{n'} p^{n'} (1-p)^{t'-n'} \; . \tag{14}$$

For odd $t_{\max}$, the probability that $\xi = -$ is denoted $p^-_{n,t} = 1 - p^+_{n,t}$. For the symmetric case, $p = 1/2$,

$$p^+_{n,t} = \frac{1}{2^{t'}} \sum_{n'=\lceil (t'-n)/2 \rceil}^{t'} \binom{t'}{n'} , \tag{15}$$

when $\lceil (t'-n)/2 \rceil > 0$ and 1 otherwise. This expression is equivalent to equation 5 in [16], which was instead expressed using $N^-_{t'}$.

The space of trajectories, i.e. of $\boldsymbol{s}_{t_{\max}}$, maps to a space of trajectories for $p^+_{n,t}$ defined on an evolving lattice in belief space. The expected reward in this case is,

$$\begin{aligned} \bar{r}_t := \langle r | N_t = n \rangle &= \mathbb{E}\left[\Theta(N_{t_{\max}} N_t) | N_t = n\right] &(16)\\ &= \max\{p^+_{n,t}, 1 - p^+_{n,t}\} &(17)\\ &= b_t \ , &(18) \end{aligned}$$

where the belief of correct report $b_t := \max\{p^+_{n,t}, 1 - p^+_{n,t}\}$. The commitment cost $\mathcal{C}^{\text{com}}_t = r_{\max} - \bar{r}_t$, then also evolves on a lattice (see fig. 3(b)). More generally, $\bar{r}_t = \Delta r b_t + r_{\text{incorrect}}$ for $\Delta r$ the difference of correct $r_{\text{correct}}$ (here 1) and incorrect $r_{\text{incorrect}}$ (here 0) rewards. Since $r_{\max} = r_{\text{correct}}$, we have $\mathcal{C}^{\text{com}}_t = \Delta r (1 - b_t)$. For $p = 1/2$ and $\Delta r = 1$, $\mathcal{C}^{\text{com}}_{t=0} = 1/2$.

The shape of $p^+_{n,t}$ is roughly sigmoidal, admitting the approximation,

$$p^+_{n,t} \approx \frac{1}{1 + \exp\left[-(at + b)n\right]} \tag{19}$$

where fitting constants $a$ and $b$ depend on $t_{\max}$. For $t_{\max} = 15$, $a = 0.03725$ and $b = 0.3557$. We demonstrate the quality of this approximation in fig. S5. Approximation error is worse at $t$ near $t_{\max}$. More than 95% of decisions times in the data we analyzed occur before 12 time steps, where the approximation error in probability is less than 0.05. A similar approximation without time dependence was presented in [16]. We nevertheless used the exact expression eq. (15) in all calculations.

## PGD implementation and fitting to relaxation after context switches

We identified the times of the context switches in the data and their type (slow-to-fast and fast-to-slow). Taking a fixed number of trials before and after each event, we averaged the decision times over the events to create two sequences of average decision times around context switches (the result is shown in fig. 4a,b). We used a uniformly weighted squared-error objective, minimized with the standard (Nelder-Mead) simplex routine in python's scientific computing library's optimization package.

## Survival probabilities over the action policy

Behavioural analyses typically focus on response time distributions. From the perspective of reinforcement learning, this is insufficient to fully characterize the behaviour of an agent. Instead, the full behaviour is given by the action policy. In this setting, a natural representation of the policy is the probability to report as a function of both the decision time *and* the environmental state (see fig. 5). These are computed from the histograms of $(N_{t^{\text{dec}}}, t^{\text{dec}})$, over trials. However, the histograms themselves do not reflect the preference of the agent to decide at a particular state and time because they are biased by the different frequencies with which the set of trajectories visit each state and time combination. While there are obviously the same number of trajectories at early and late times, they distribute over many more states at later times and so each state at later times is visited less on average than states at earlier times. We can remove this bias by transforming the data ensemble to the ensemble of two random variables: the state conditioned on time $(N_t|t)$, and the event that $t = t^{\mathrm{dec}}$. Conditioning this ensemble on the state gives $P(t = t^{\mathrm{dec}}|N_t, t) = p(N_t, t = t^{\mathrm{dec}}|t)/p(N_t|t)$. To reduce estimator variance, we focus on the corresponding survival function, $P(t < t^{\mathrm{dec}}|N_t, t)$. So, $P(t < t^{\mathrm{dec}}|N_t, t) = 1$ when $t = 0$ and decays to 0 as $t$ and $|N_t|$ increase. Unlike the unconditioned histograms, these survival probabilities vary much more smoothly over state and time. This justifies the use of the interpolated representations displayed in fig. 5b-e. Note that to simplify the analysis, we have binned decision times by the 200 ms time step between token jumps. This is justified by the small deviations from uniformity of decision times modulo the time step shown in fig. S11.

### Episodic decision-making and dynamic programming solutions of value iteration

We generalize the mathematical notation and description of an existing AR-RL formulation and dynamic programming solution of the random dots task [7], a binary perceptual evidence accumulation task extensively studied in neuroscience. To align notation with convention in reinforcement learning theory, exceptionally here $s$ denotes the belief state variable, ie. a representation of the task state sufficient to make the decision (e.g. the tokens difference, $N_t$, in the case of the tokens task). We connect this extended formulation to account for a dynamic deliberation cost. We write it in discrete time, though the continuous time version is equally tractable.

The problem is defined by a recursive optimality equation for the value function $V(s|t)$ in which the highest of the action values, $Q(s, a|t)$, is selected. We formalize the non-stationarity within episodes by conditioning on the trial time, $t$, where $t = 0$ is the trial start time. $Q(s, a|t)$ is the action-value function of average-reward reinforcement learning [11], i.e. the expected sum of future reward deviations from the average when selecting action $a$ when in state $s$, at possible decision time $t$ within a trial, and then following a given action policy $\pi$ thereafter. The action set for these binary decision tasks consists of *report left* $(-)$, *report right* $(+)$, and *wait*. When *wait* is selected, time increments and beliefs are updated with new evidence. We use a decision-time conditioned, expected trial reward function, $r(s, a|t)$ with $a = \pm$, that denotes the reward expected to be received at the end of the trial after having reported $\pm$ in state $s$ at time $t$ during the trial. Note that $r(s, a|t)$ can be defined in terms of a conventional reward function $r(s, a)$ if the reported action, decision time, and current time are stored as an auxiliary state variable so they can be used to determine the non-zero reward entries at the end of the trial.

The average-reward formulation of $Q(s, a|t)$ naturally narrows the problem onto determining decisions within only a single episode of the task. To see this, we pull out the contribution of the current trial,

$$Q(s, a|t) = \mathbb{E}^{\pi}\left[\sum_{t'=t}^{T} R_t - \rho \,\middle|\, S_t = s, A_t = a\right] + V(s|T+1) \tag{20}$$

where $T$ is the (possibly stochastic) trial end time and $V(s|T+1)$ is the state value at the start of the following trial, which does not depend on $s_t$ and $a_t$ for independently sampled trials. Following conventional reinforcement learning notation, the expectation $\mathbb{E}^{\pi}$ is over all randomness conditioned on following the policy, $\pi$, which itself could be stochastic [11]. When trials are identically and independently sampled, the state at the trial start is the same for all trials and denoted $s_0$ with value $V_0$. Thus, the value at the start of the trial $V(s|t=0) = V(s|T+1) = V_0$ equals that at the start of the next trial and so, by construction, the expected trial return (total trial rewards minus trial costs) must vanish (we will show this explicitly below). Note that the value shift invariance of eq. (20) can be fixed so that $V_0 = 0$.

The *optimality equation* for $V(s|t)$ arises from a greedy action policy over $Q(s,a|t)$: it selects the action of the largest of $Q(s,-|t)$, $Q(s,+|t)$, and $Q(s,wait|t)$. The value expression for the wait-action is incremental, and so depends on the value at the next time step. In contrast, expression for the two reporting actions integrates over the remainder of the trial since no further decision is made and so depends on the value at the start of the following trial. The resulting optimality equation for the value function $V(s|t)$ is then

$$\begin{aligned} V(s|t) &= \max_a Q(s,a|t) \ , \\ Q(s,\pm|t) &= r(s,\pm|t) - \sum_{t'=t+1}^{T} c_{t'} + V(s|t=T+1) \ , \\ Q(s,wait|t) &= -c_t + \mathbb{E}_{s_{t+1}|s}\left[V(s_{t+1}|t+1)\right] \ , \\ V(s|t=0) &= V(s|t=T+1) \ . \end{aligned} \tag{21}$$

Here, $t = 0, 1, \ldots, t_{\max}$ within the current trial and $t = T+1, T+2 \ldots$ in the following trial, with $t_{\max}$ the latest possible decision time in a trial, and $T = T(t)$ the decision-time dependent trial duration. For inter-trial interval $T_{\mathrm{ITI}}$, $T$ satisfies $T_{\mathrm{ITI}} \leq T \leq t_{\max} + T_{\mathrm{ITI}}$. $c_t$ is the cost rate at time $t$. The second term in $Q(s,wait|t)$ uses the notation $\mathbb{E}_{x|y}[z]$, i.e. the expectation of $z$ with respect to $p(x|y)$. The last line in eq. (21) is the self-consistency criterion imposed by the AR-RL formulation, which demands that the expected value at the beginning of the trial be the expected value at the beginning of the following trial. The greedy policy then gives a single decision time for each state trajectory as the first time when $Q(s,-|t) > Q(s,wait|t)$ or $Q(s,+|t) > Q(s,wait|t)$, with the reporting action determined by which of $Q(s,-|t)$ and $Q(s,+|t)$ is larger. For given $c_t$, dynamic programming provides a solution to eq. (21) [7] by recursively solving for $V(s|t)$ by back-iterating in time from the end of the trial. For most relevant tasks, to never report is always sub-optimal, so the value at $t = t_{\max}$ is set by the best of the two reporting ($\pm$) actions, which do not have a recursive dependence on the value and so can seed the recursion.

We now interpret this general formulation in terms of opportunity costs. For the choice of a static opportunity cost rate of time, $c_t = \rho$. This is the AR-RL case. As in [7], a constant auxiliary deliberation cost rate, $c$, incurred only up to decision time can be added, $c_t = \rho + c\Theta(t^{\mathrm{dec}} - t)$. Of course, $\rho$ is unknown *a priori*. For this solution method, its value can be found by exploiting the self-consistency constraint, $V(s|t=0) = V(s|t=T+1)$. This dependence can be seen formally by taking the action value eq. (20), choosing $a$ according to $\pi$ to obtain the state value, $V(s|t)$, and evaluating it for $t = 0$,

$$V(s|t=0) = \mathbb{E}_{t^{\text{dec}}}\left[\sum_{t=0}^{T} R_t - \rho\right] + V(s|t = T+1) \tag{22}$$

$$= \mathbb{E}_{t^{\text{dec}}}\left[r(t^{\text{dec}}) - \rho T(t^{\text{dec}})\right] + V(s|t = T+1) \tag{23}$$

$$= \bar{R} - \rho\bar{T} + V(s|t = T+1)\,. \tag{24}$$

Here, $\bar{R} = \mathbb{E}_{t^{\text{dec}}}\left[r(t^{\text{dec}})\right]$ and $\bar{T} = \mathbb{E}_{t^{\text{dec}}}\left[T(t^{\text{dec}})\right]$ denotes the expectations over the trial ensemble that, when given the state sequence, transforms to an average over $t^{\text{dec}}$, the trial decision time, defined as when $V(s|t)$ achieves its maximum on the state sequence, $(s_0, \ldots, s_{t_{\max}})$. The expected trial reward function, $r(t) := \max_{a\in\{-,+\}} r(s, a|t)$ is the expected trial reward for deciding at $t$. Imposing the self-consistency constraint on eq. (24) recovers the definition $\rho = \bar{R}/\bar{T}$.

## Asymmetric switching cost model

Here, we present the model component that accounts for the asymmetric relaxation timescales after context switches. The basic assumption is that tracking a signal at a higher temporal resolution should be more cognitively costly, so that adapting from faster to slower environments should happen more quickly than the reverse, so as to not pay this cost unnecessarily. We now develop this idea formally (see fig. S4).

Let $T_{\text{track}}$ and $T_{\text{sys}}$ be the timescale of tracking and of the tracked system, respectively. One way to interpret the mismatch ratio, $T_{\text{sys}}/T_{\text{track}}$, is via an attentional cost rate, $q$. This rate should decay with $T_{\text{track}}$: the slower the timescale of tracking, the lower the cognitive cost. For simplicity, we set $q = 1/T_{\text{track}}$ (fig. S4a). Integrating this cost rate over a characteristic time of the system is then the tracking cost, $Q = qT_{\text{sys}} = T_{\text{sys}}/T_{\text{track}}$, which is also the mismatch ratio. We propose that $Q$ enters the algorithm via a scale factor on the integration time of the reward filter for $\hat{\rho}_k^{\tau_{\text{context}}}$, $\tau_{\text{context}}$. We redefine $\tau_{\text{context}}$ as

$$\tau_{\text{context}} \leftarrow \frac{\tau_{\text{context}}}{1+Q^{\nu}}\,, \tag{25}$$

where $\nu$ is a sensitivity parameter that captures the strength of the nonlinear sensitivity of the speed up (for $\nu > 1$) or slow down (for $\nu < 1$) in adaptation with the tracking cost, $Q$ (fig. S4a shows how this timescale varies over $Q$ for three values of $\nu$). A natural choice for $T_{\text{sys}}$ is $T_k$, the trial duration. For $T_{\text{track}}$, we introduce the filtered estimate of the trial duration, $\hat{T}_k^{\tau_{\text{context}}}$ (computed using the same simple low-pass filter *c.f.* eq. (8)). Thus, the tracking timescale adapts to the system timescale. As a result of how $\tau_{\text{context}}$ is lowered by $Q$ for $\nu > 1$, this adaptation is faster in the fast-to-slow transition relative to the slow-to-fast transition.

## Prediction for asymmetric rewards

Given a payoff matrix, $\boldsymbol{R} = (r_{s,a})$, where $r_{s,a}$ is the reward for reporting $a \in \{-,+\}$ in the trial realization leading to $s$, here the sign of $N_{t_{\max}}$, and the probability that the rightward choice is correct, $p^+_{n,t}$, the expected reward for the two reporting actions in a trial is given by the matrix equation

$$\left[\langle r|a=+,n,t\rangle \;\; \langle r|a=-,n,t\rangle\right] = \left[p^+_{n,t} \;\; 1-p^+_{n,t}\right] \begin{bmatrix} r_{++} & r_{+-} \\ r_{-+} & r_{--} \end{bmatrix} .$$

Here, the corresponding reported choice is $a^* = \text{argmax}_{a\in\{-,+\}}\langle r|a,n,t\rangle$. In this paper and in all existing tokens tasks, $\boldsymbol{R}$ was the identity matrix. In this case, and for all cases where $\boldsymbol{R}$ is a symmetric matrix, $\boldsymbol{R} = \boldsymbol{R}^\top$, an equivalent decision rule is to decide based on the sign of $N_t$. When $\boldsymbol{R}$ is not symmetric, however, this is no longer a valid substitute. Asymmetry can be introduced through the actions and the states.

Using an additional parameter $\gamma$, we introduce asymmetry via a bias for $+$ actions that leaves the total reward unchanged by replacing the payoff matrix with

$$\boldsymbol{R}_{\text{asym}} = \begin{bmatrix} r_{++}(1+\gamma) & r_{+-}(1-\gamma) \\ r_{-+}(1+\gamma) & r_{--}(1-\gamma) \end{bmatrix} ,$$

The result for $\gamma = -0.6, 0,$ and $0.6$ is shown in fig. S10. For $\gamma > 0$ the decision boundary for $a = +$ shifts up proportional to $\gamma$. For $\gamma < 0$ the decision boundary for $a = -$ shifts down proportional to $-\gamma$. The explanation is that the components are set and exchange where the decision is exchanged, $N_t = 0$ for the symmetric case. This changes to $N_t \propto \pm\gamma$ for the asymmetric $\gamma \neq 0$ case.

### Comparing reward rates and slopes of urgency

Reference [17] parametrize urgency with the saturation value, $u_\infty$, and the half-maximum, $\tau_{1/2}$. The initial slope is given by their ratio. We used the context-conditioned values published in Table 1 in [17] for the $n = 70$ (no 90° control) dataset. The context-conditioned reward rates, $\rho_\alpha$, are computed as the accuracy $\langle R\rangle_{|\alpha}$ divided by the average trial time, $\langle T\rangle_{|\alpha}$ for choice number $\alpha \in \{2,4\}$ as context. We computed $\langle R\rangle_{|\alpha=2} = 0.71$ and $\langle R\rangle_{|\alpha=4} = 0.49$. The trial time is the sum of the response time, the added time penalty if incorrect, and the inter-trial interval. We computed the response times $t_{\text{response},\alpha=2} = 0.527$ and $t_{\text{response},\alpha=4} = 0.725$. While the dataset contains the response times, it does not have the latter two. The time penalty was on the order of 1 second, as was the time penalty [61]. Under those estimates, the reward rates are $\rho_{\alpha=2} = 0.40$ and $\rho_{\alpha=4} = 0.22$. The ratio between slopes is 1.8 and the ratio of reward rates was 2.3 giving an error of about 20%.


### ACKNOWLEDGMENTS

We would like to acknowledge helpful discussions with Jan Drugowitsch, Becket Ebitz, and Paul Masset, and to Anne Churchland for sharing data from [17]. MPT acknowledges support from IVADO via their postdoctoral fellowship award. PC acknowledges support from NSERC Discovery Grant (RGPIN-2016-05245). GL acknowledges support from FRQS Research Scholar Award, Junior 1 (LAJGU0401-253188), NSERC Discovery Grant (RGPIN-

2018-04821), and the Canada CIFAR AI Chair program.

---


[1] David I Green, "Pain-Cost and Opportunity-Cost," The Quarterly Journal of Economics **8**, 218–229 (1894).

[2] Vektor Dewanto, George Dunn, Ali Eshragh, Marcus Gallagher, and Fred Roosta, "Average-reward model-free reinforcement learning: a systematic review and literature mapping," arXiv:2010.08920 [cs.LG].

[3] Nathaniel D Daw and David S Touretzky, "Long-term reward prediction in TD models of the dopamine system," Neural computation **14**, 2567–2583 (2002).

[4] Lindsay E Hunter and Nathaniel D Daw, "Context-sensitive valuation and learning," Current Opinion in Behavioral Sciences **41**, 122–127 (2021).

[5] Nils Kolling and Thomas Akam, "(Reinforcement?) Learning to forage optimally," Current Opinion in Neurobiology **46**, 162–169 (2017).

[6] Yael Niv, Nathaniel D Daw, and Peter Dayan, "How fast to work : Response vigor , motivation and tonic dopamine," in *Neural Information Processing Systems* (2005).

[7] Jan Drugowitsch, Rubén Moreno-Bote, Anne K Churchland, Michael N Shadlen, and Alexandre Pouget, "The Cost of Accumulating Evidence in Perceptual Decision Making," The Journal of Neuroscience **32**, 3612 LP – 3628 (2012).

[8] A Ross Otto and Nathaniel D Daw, "The opportunity cost of time modulates cognitive effort," Neuropsychologia **123**, 92–105 (2019).

[9] A Ross Otto and Eliana Vassena, "It's all relative: Reward-induced cognitive control modulation depends on context." Journal of Experimental Psychology: General **150**, 306–313 (2021).

[10] Germain Lefebvre, Aurélien Nioche, Sacha Bourgeois-gironde, and Stefano Palminteri, "Contrasting temporal difference and opportunity cost reinforcement learning in an empirical money-emergence paradigm," Proceedings of the National Academy of Sciences **115**, E11446 LP – E11454 (2018).

[11] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction, 2nd ed.*, Adaptive computation and machine learning. (The MIT Press, Cambridge, MA, US, 2018) pp. xxii, 526–xxii, 526.

[12] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup, "Towards Continual Reinforcement Learning: A Review and Perspectives," arXiv:2012.13490 [cs.LG].

[13] Roger Ratcliff, "A theory of memory retrieval." Psychological Review **85**, 59–108 (1978).

[14] Gaurav Malhotra, David S Leslie, Casimir J H Ludwig, and Rafal Bogacz, "Time-varying decision boundaries : insights from optimality analysis," Psychon Bull Rev **25**, 971–996 (2018).

[15] Jochen Ditterich, "Evidence for time-variant decision making," European Journal of Neuroscience **24**, 3628–3641 (2006).

[16] Paul Cisek, Geneviève Aude Puskas, and Stephany El-Murr, "Decisions in Changing Conditions: The Urgency-Gating Model," The Journal of Neuroscience **29**, 11560 LP – 11571 (2009).

[17] Anne K Churchland, Roozbeh Kiani, and Michael N Shadlen, "Decision-making with multiple alternatives," Nature Neuroscience **11**, 693–702 (2008).

[18] David Thura and Paul Cisek, "Deliberation and Commitment in the Premotor and Primary Motor Cortex during Dynamic Decision Making," Neuron **81**, 1401–1416 (2014).

[19] David Thura, Ignasi Cos, Jessica Trung, and Paul Cisek, "Context-Dependent Urgency Influences Speed–Accuracy Trade-Offs in Decision-Making and Movement Execution," The Journal of Neuroscience **34**, 16442 LP – 16454 (2014).

[20] David Thura, Jean-François Cabana, Albert Feghaly, and Paul Cisek, "Unified neural dynamics of decisions and actions in the cerebral cortex and basal ganglia," bioRxiv (2020), 10.1101/2020.10.22.350280.

[21] David Thura and Paul Cisek, "The Basal Ganglia Do Not Select Reach Targets but Control the Urgency of Commitment," Neuron **95**, 1160–1170.e5 (2017).

[22] Peter Janssen and Michael N Shadlen, "A representation of the hazard rate of elapsed time in macaque area LIP," Nature Neuroscience **8**, 234–241 (2005).

[23] Satohiro Tajima, Jan Drugowitsch, and Alexandre Pouget, "Optimal policy for value-based decision-making," Nature Communications **7**, 12400 (2016).

[24] Anton Schwartz, "A Reinforcement Learning Method for Maximizing Undiscounted Rewards," in *International Conference on Machine Learning*, Vol. 0 (1993).

[25] Yael Niv, Nathaniel D Daw, Daphna Joel, and Peter Dayan, "Tonic dopamine: opportunity costs and the control of response vigor," Psychopharmacology **191**, 507–520 (2007).

[26] Sara M Constantino and Nathaniel D Daw, "Learning the opportunity cost of time in a patch-foraging task," Cogn Affect Behav Neurosci. **15**, 837 (2015).

[27] Benjamin Y Hayden and Yael Niv, "The case against economic values in the orbitofrontal cortex (or anywhere else in the brain)," PsyArXiv 10.31234/osf.io/7hgup.

[28] Nathaniel D Daw, "Advanced Reinforcement Learning," in *Neuroeconomics*, edited by Paul W Glimcher and Ernst B T Neuroeconomics (Second Edition) Fehr (Academic Press, San Diego, 2014) 2nd ed., Chap. 16, pp. 299–320.

[29] D. Thura. Personal communication.

[30] Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum, "One and Done? Optimal Decisions From Very Few Samples," Cognitive Science **38**, 599–637 (2014).

[31] Single subject behavioural data shared by Thomas Thierry.

[32] Surya Ganguli, James W Bisley, Jamie D Roitman, Michael N Shadlen, Michael E Goldberg, and Kenneth D Miller, "One-Dimensional Dynamics of Attention and Decision Making in LIP," Neuron **58**, 15–25 (2008).

[33] David Thura and Paul Cisek, "Modulation of Premotor and Primary Motor Cortical Activity during Volitional Adjustments of Speed-Accuracy Trade-Offs," The Journal of Neuroscience **36**, 938 – 956 (2016).

[34] Kiyohito Iigaya, Yashar Ahmadian, Leo P Sugrue, Greg S Corrado, Yonatan Loewenstein, William T Newsome, and Stefano Fusi, "Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales," Nature Communications **10**, 1466 (2019).

[35] Long Ding and Joshua I. Gold, "The Basal Ganglia's Contributions to Perceptual Decision Making," Neuron **79**, 640–649 (2013).

[36] Kong-Fatt Wong and Xiao-Jing Wang, "A Recurrent Network Mechanism of Time Integration in Perceptual Decisions," The Journal of Neuroscience **26**, 1314 – 1328 (2006).

[37] Alex Roxin and Anders Ledberg, "Neurobiological Models of Two-Choice Decision Making Can Be Reduced to a One-Dimensional Nonlinear Diffusion Equation," PLOS Computational Biology **4**, e1000046 (2008).

[38] David Meder, Nils Kolling, Lennart Verhagen, Marco K Wittmann, Jacqueline Scholl, Kristoffer H Madsen, Oliver J Hulme, Timothy E J Behrens, and Matthew F S Rushworth, "Simultaneous representation of a spectrum of dynamically changing value estimates during decision

making," Nature Communications **8** (2017), 10.1038/s41467-017-02169-w.

[39] "Predictive Representations in Hippocampal and Prefrontal Hierarchies," .

[40] Jan Zimmermann, Paul W Glimcher, and Kenway Louie, "Multiple timescales of normalized value coding underlie adaptive choice behavior," Nature Communications **9**, 3206 (2018).

[41] HyungGoo R Kim, Athar N Malik, John G Mikhael, Pol Bech, Iku Tsutsui-Kimura, Fangmiao Sun, Yajun Zhang, Yulong Li, Mitsuko Watabe-Uchida, Samuel J Gershman, and Naoshige Uchida, "A Unified Framework for Dopamine Signals across Timescales," Cell **183**, 1600–1616.e25 (2020).

[42] Paul Masset, Athar N. Malik, HyungGoo R. Kim, Pol Bech, and Naoshige Uchida, "A diversity of discounting horizons explains ramping diversity in dopaminergic neurons," in *COSYNE Abstracts* (2021).

[43] Angela J Langdon and Nathaniel D Daw, "Beyond the Average View of Dopamine," Trends in Cognitive Sciences **24**, 499–501 (2020).

[44] John G Mikhael and Samuel J Gershman, "Adapting the flow of time with dopamine," Journal of Neurophysiology **121**, 1748–1760 (2019).

[45] Ido Toren, Kristoffer C Aberg, and Rony Paz, "Prediction errors bidirectionally bias time perception," Nature Neuroscience **23**, 1198–1202 (2020).

[46] Lars Hunger, X Arvind Kumar, and X Robert Schmidt, "Abundance Compensates Kinetics : Similar Effect of Dopamine Signals on D1 and D2 Receptor Populations," Journal of Neuroscience **40**, 2868–2881.

[47] Julia Cox and Ilana B Witten, "Striatal circuits for reward learning and decision-making," Nature Reviews Neuroscience **20**, 482–494 (2019).

[48] Helen N Schwerdt, Hideki Shimazu, Ken-ichi Amemori, Satoko Amemori, Patrick L Tierney, Daniel J Gibson, Simon Hong, Tomoko Yoshida, Robert Langer, Michael J Cima, and Ann M Graybiel, "Long-term dopamine neurochemical monitoring in primates," Proceedings of the National Academy of Sciences **114**, 13260 LP – 13265 (2017).

[49] Tommaso Patriarchi, Jounhong Ryan Cho, Katharina Merten, Mark W Howe, Aaron Marley, Wei-hong Xiong, Robert W Folk, Gerard Joey Broussard, Ruqiang Liang, Min Jee Jang, Haining Zhong, Daniel Dombeck, Mark Von Zastrow, Axel Nimmerjahn, Viviana Gradinaru, John T Williams, and Lin Tian, "Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors," Science **4422** (2018), 10.1126/science.aat4422.

[50] Matthew A Carland, David Thura, and Paul Cisek, "The Urge to Decide and Act: Implications for Brain Function and Dysfunction," The Neuroscientist **25**, 491–511 (2019).

[51] Samuel J Gershman and Naoshige Uchida, "Believing in dopamine," Nature Reviews Neuroscience **20**, 703–714 (2019).

[52] Andrew Westbrook and Todd S Braver, "Dopamine Does Double Duty in Motivating Cognitive Effort," Neuron **91**, 708 (2016).

[53] Pablo Tano, Peter Dayan, and Alexandre Pouget, "A Local Temporal Difference Code for Distributional Reinforcement Learning," in *Advances in Neural Information Processing Systems*, Vol. 33, edited by H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (Curran Associates, Inc., 2020) pp. 13662–13673.

[54] William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle, "Hyperbolic Discounting and Learning over Multiple Horizons," arXiv:1902.06865 [stat.ML].

[55] I Momennejad, E M Russek, J H Cheong, M M Botvinick, N D Daw, and S J Gershman, "The successor representation in human reinforcement learning," Nature Human Behaviour **1**, 680–692 (2017).

[56] Personal communication, Thomas Thierry.

[57] Stefano Palminteri and Maël Lebreton, "Context-dependent outcome encoding in human reinforcement learning," Current Opinion in Behavioral Sciences **41**, 144–151 (2021).

[58] Ernest S Davis and Gary F Marcus, "Computational limits don't fully explain human cognitive limitations," Behavioral and Brain Sciences **43**, e7 (2020).

[59] Arman Abrahamyan, Laura Luz Silva, Steven C Dakin, Matteo Carandini, and Justin L Gardner, "Adaptable history biases in human perceptual decisions," Proceedings of the National Academy of Sciences **113**, E3548 LP – E3557 (2016).

[60] Zhijian Wang, Bin Xu, and Hai-Jun Zhou, "Social cycling and conditional responses in the Rock-Paper-Scissors game," Scientific Reports **4**, 5830 (2014).

[61] A. Churchland. Personal communication.

1049
# Supplemental Materials

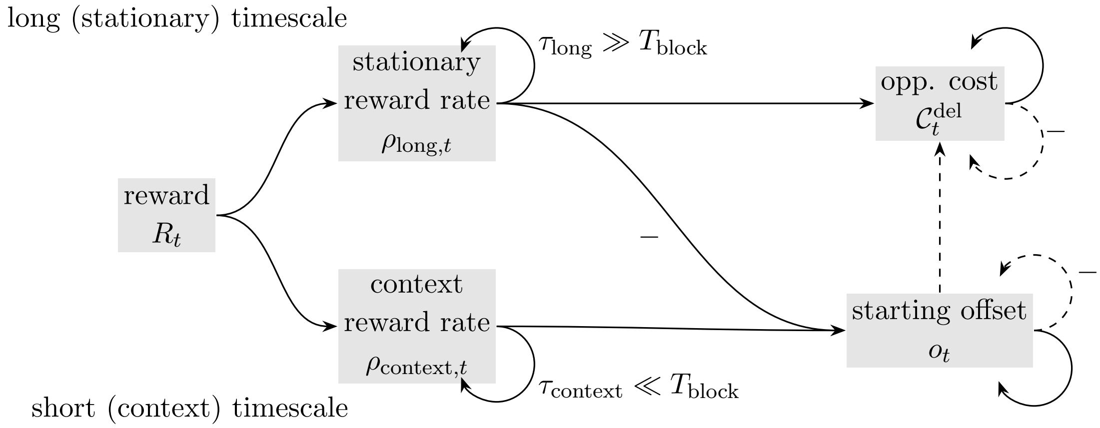



Figure S1. *Reward filtering scheme for online computation of within-trial opportunity cost.* With $t$ denoting absolute time, the reward sequence, $R_t$, is integrated on both a stationary ($\tau_{\text{long}}$) and context ($\tau_{\text{context}}$) filtering timescale to produce estimates of the stationary and context-specific reward rates, respectively. These are large and small, respectively, relative to the average context switching timescale, $T_{\text{block}}$. The estimate of the context-specific offset, $o_t$ is computed by time-integrating the difference of these two estimates. In this filtering, when a trial terminates, the effective operation is that $\mathcal{C}_t^{\text{del}}$ is set to $o_t$, and the latter is zeroed. Thus, the opportunity cost starts at this offset and then integrates $\rho_{\text{long}}$, $\mathcal{C}_{t,k}^{\text{del}} = o_{T_{k-1},k-1} + \rho_{\text{long},k-1}t$, where $o_{T_{k-1},k-1} = (\rho_{\text{context},k-1} - \rho_{\text{long},k-1})T_{k-1}$. Notes on the computational graph: Arrows pass the value at each time step (dashed arrows only pass the value when a trial terminates). Links annotated with '−' multiply the passed quantity by −1.

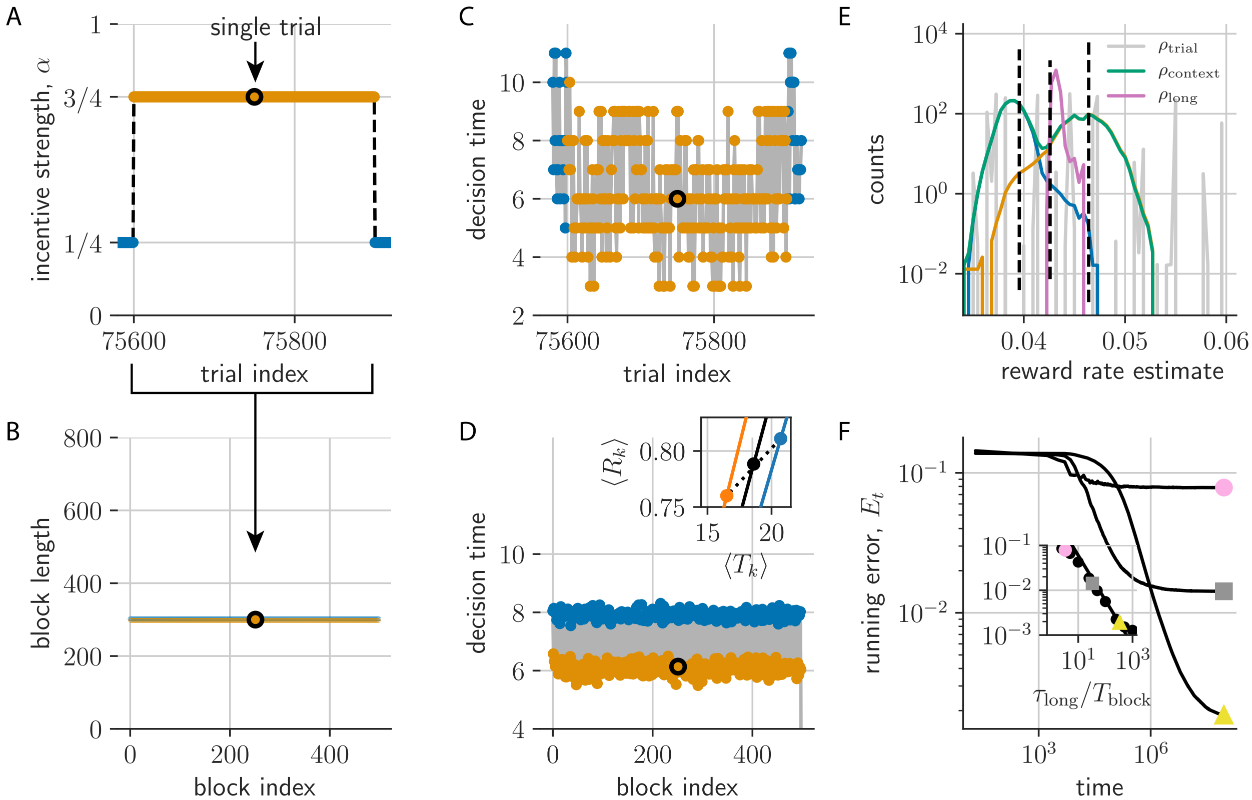



Figure S2. *PGD agent plays the tokens task with periodic $\alpha$-dynamics.* (a) Trials are grouped into alternating trial blocks of constant $\alpha$ (fast (orange) and slow (blue) conditions). (b) Here, trial block durations are constant over the experiment. (c) Decision times over the trials from (a) distribute widely, but relax after context switches. (d) Block-averaged decision times remain stationary. Inset shows the context-conditioned trial-averaged reward $\langle R_k \rangle$ and trial duration $\langle T_k \rangle$ (orange and blue dots; black is unconditioned average; $\langle \cdot \rangle$ denotes the trial ensemble average). Lines pass through the origin (slope given by the respective reward rate). (e) Distribution of estimates have lower variance than the trial reward rates, $\rho^{\text{trial}}$ (gray). The conditioned averages of $\hat{\rho}_k^{\tau_{\text{context}}}$ shown as blue and orange. (f) The relative error in estimating $\rho$, $E_t = \frac{1}{t}\sum_k^t |\hat{\rho}_k^{\tau_{\text{long}}} - \rho|/\rho$, for $\tau_{\text{long}} = 10^3$(circle), $10^4$(square), $10^5$(triangle). Inset shows that $E_{T_{\text{exp}}} \propto (\tau_{\text{long}}/T_{\text{block}})^{-1}$ over a grid of $\tau_{\text{long}}$ and $T_{\text{block}}$ as expected (black line).

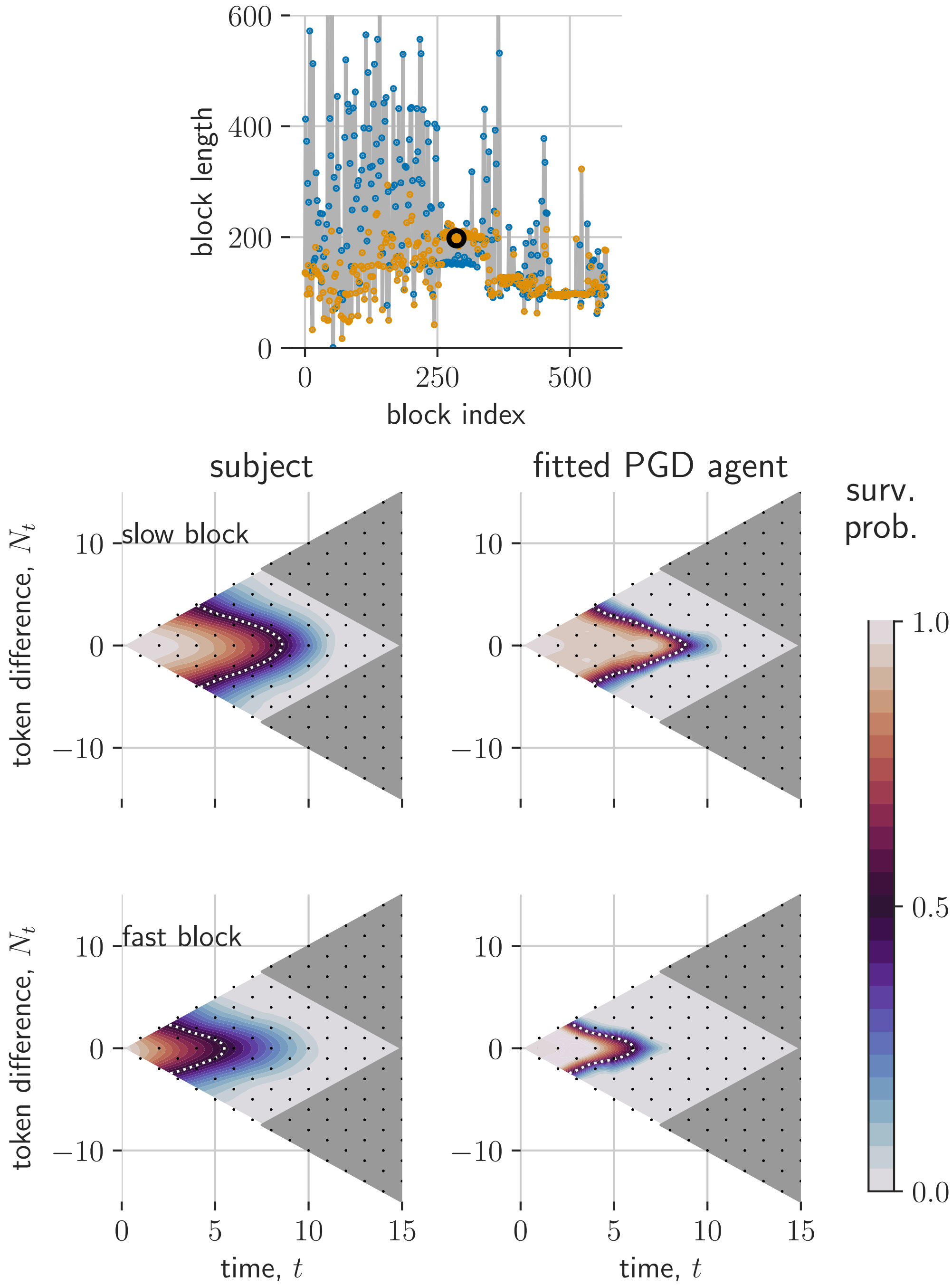



Figure S3. *Comparison of PGD and NHP in non-stationary $\alpha$ dynamics from [19]: Subject 2.* Same as fig. 5.

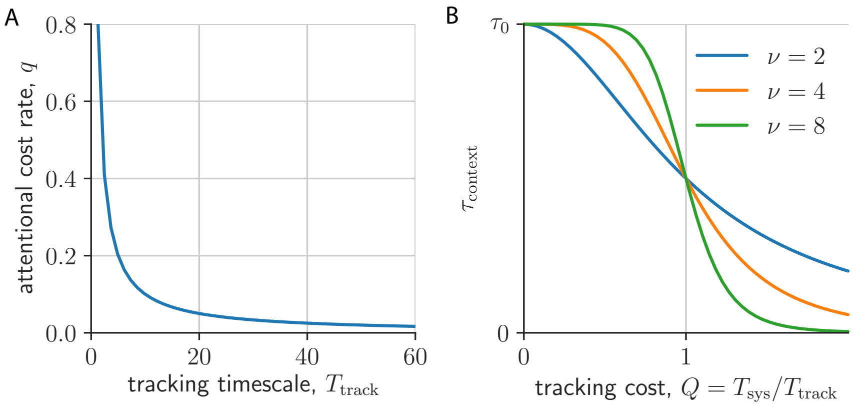



Figure S4. *Asymmetric switching cost model.* (a) Attentional cost rate, $q$, is set to be inversely proportional to tracking timescale, $T_{\text{track}}$. (b) Filtering timescale $\tau_{\text{context}}$ is scaled down with tracking cost, $Q = T_{\text{sys}}/T_{\text{track}}$ from a base timescale, here denoted $\tau_0$ (shown for three values of sensitivity $\nu = 2, 4, 8$).

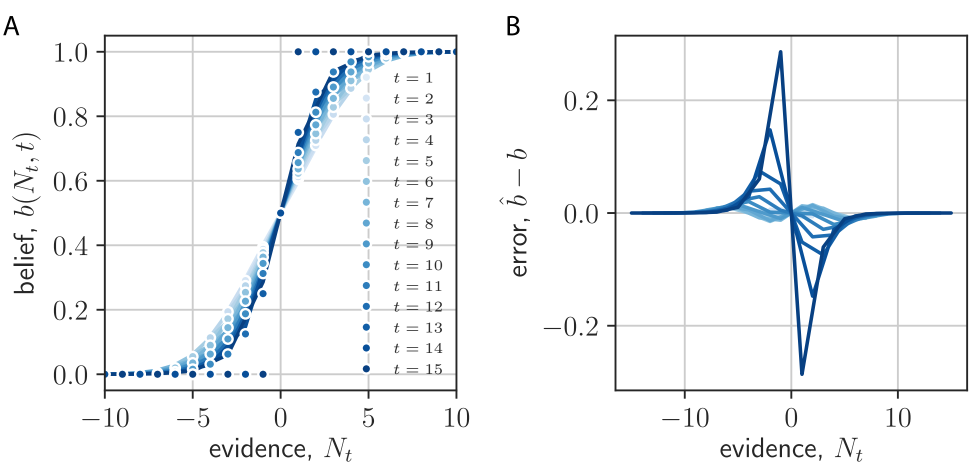



Figure S5. *sigmoidal approximation to expected reward.* (a) the approximation explained in Methods: State-conditioned expected trial reward, for different dec,[p]ision times. (b) The error in the approximation for different decision times.

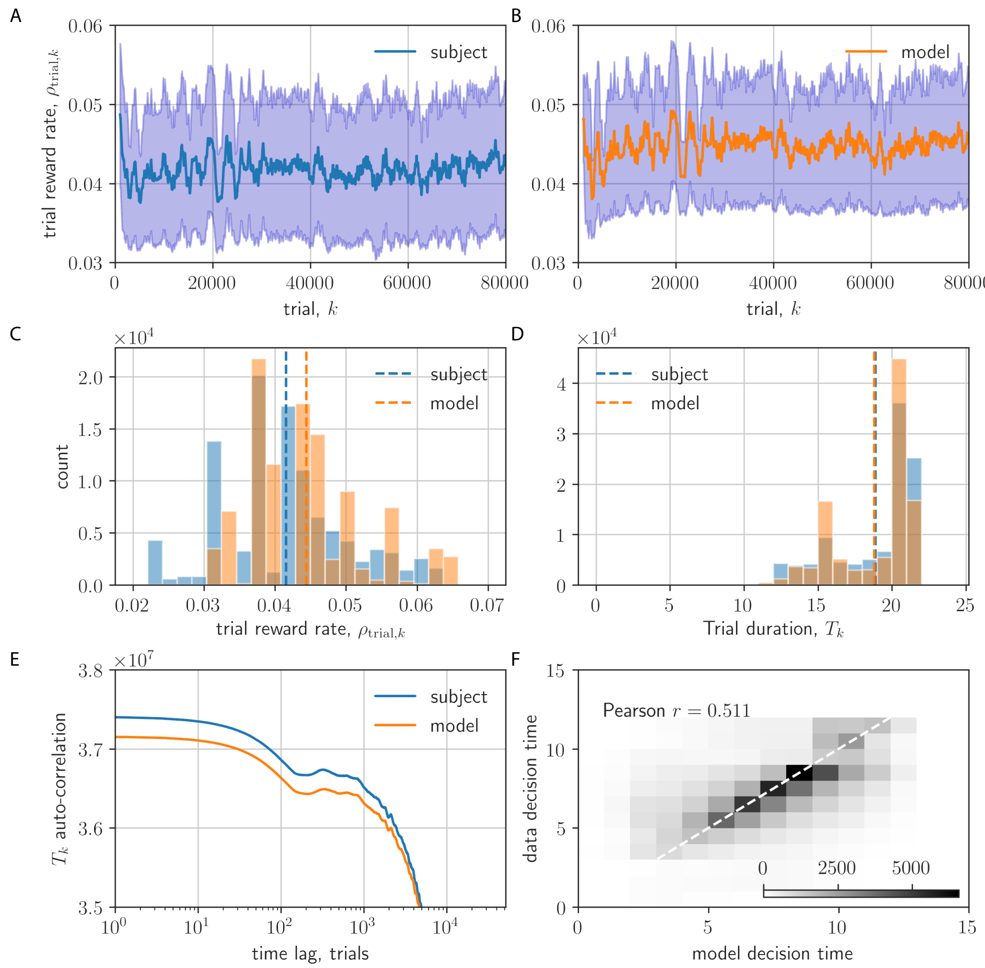



Figure S6. *Model validation on behavioural statistics from [19].* (a,b) Running average (last 1000 trial) of trial reward rate $\rho_k^{\mathrm{trial}}$. (c,d) Histograms of trial reward rate, $\rho_k^{\mathrm{trial}}$ (c) and trial duration, $T_k$ (d). (e) Auto-correlation function of trial duration. (f) Data vs. model decision time (gray-scale is count; white dashed line is perfect correlation; actual Pearson correlation is shown)

.

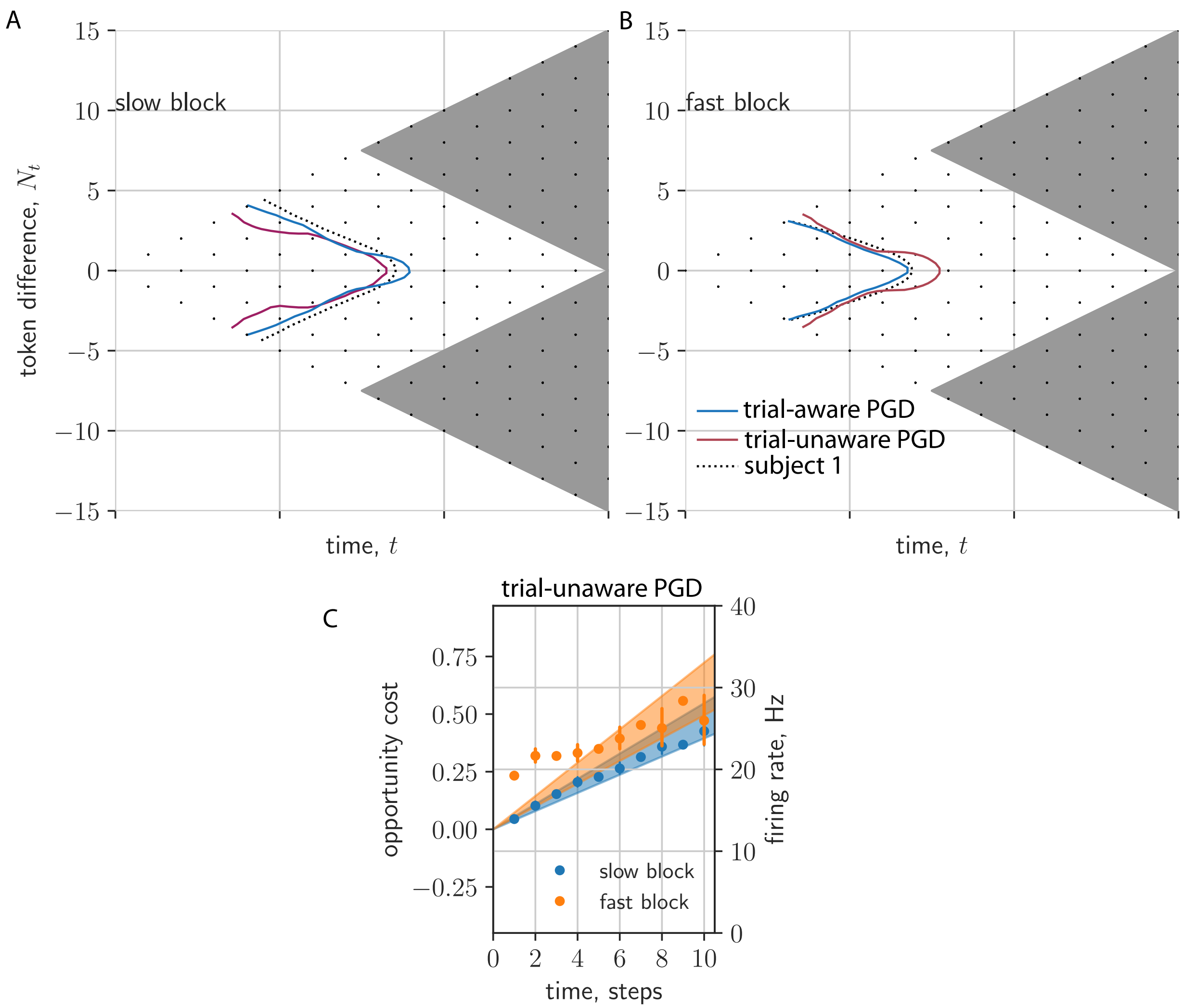
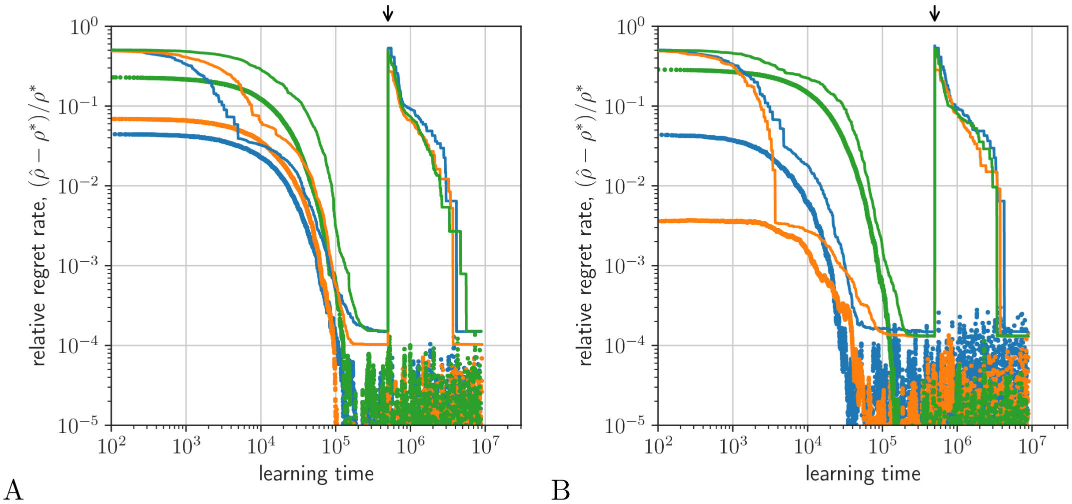



Figure S7. *Comparison of trial-aware and trial-unaware results.* (a,b) 1/2-Survival probability contours for subject 1 (dashed), trial-aware PGD (blue), and trial-unaware PGD (red) for slow (a) and fast (b) context-conditioned data. (c) Opportunity cost for trial-unaware PGD (compare with fig. 2b). Opportunity cost range adjusted here such that data within standard error of trial-unaware PGD model prediction for slow block (blue).




A B

Figure S8. *Comparison of PGD and AR-RL learning on a patch leaving task.* Performance is defined as relative regret rate, $(\hat{\rho} - \rho^*)/\rho^*$ (PGD (dots); AR-RL (lines)). (a) Performance over different sizes of the state vector ($d = 100$ (blue), 200 (orange), 300 (green)). (b) Performance over different learning rates (parametrized by integration time constant, $\tau = 1 \times 10^4$ (blue), $2 \times 10^4$ (orange), $3 \times 10^4$ (green)). (parameters: $\lambda = 1/5$; $r_{\max}$ sampled uniformly on $[0, 1]$). A random state label permutation is made at the time indicated by the black arrow. Values were initialized at $-1$.

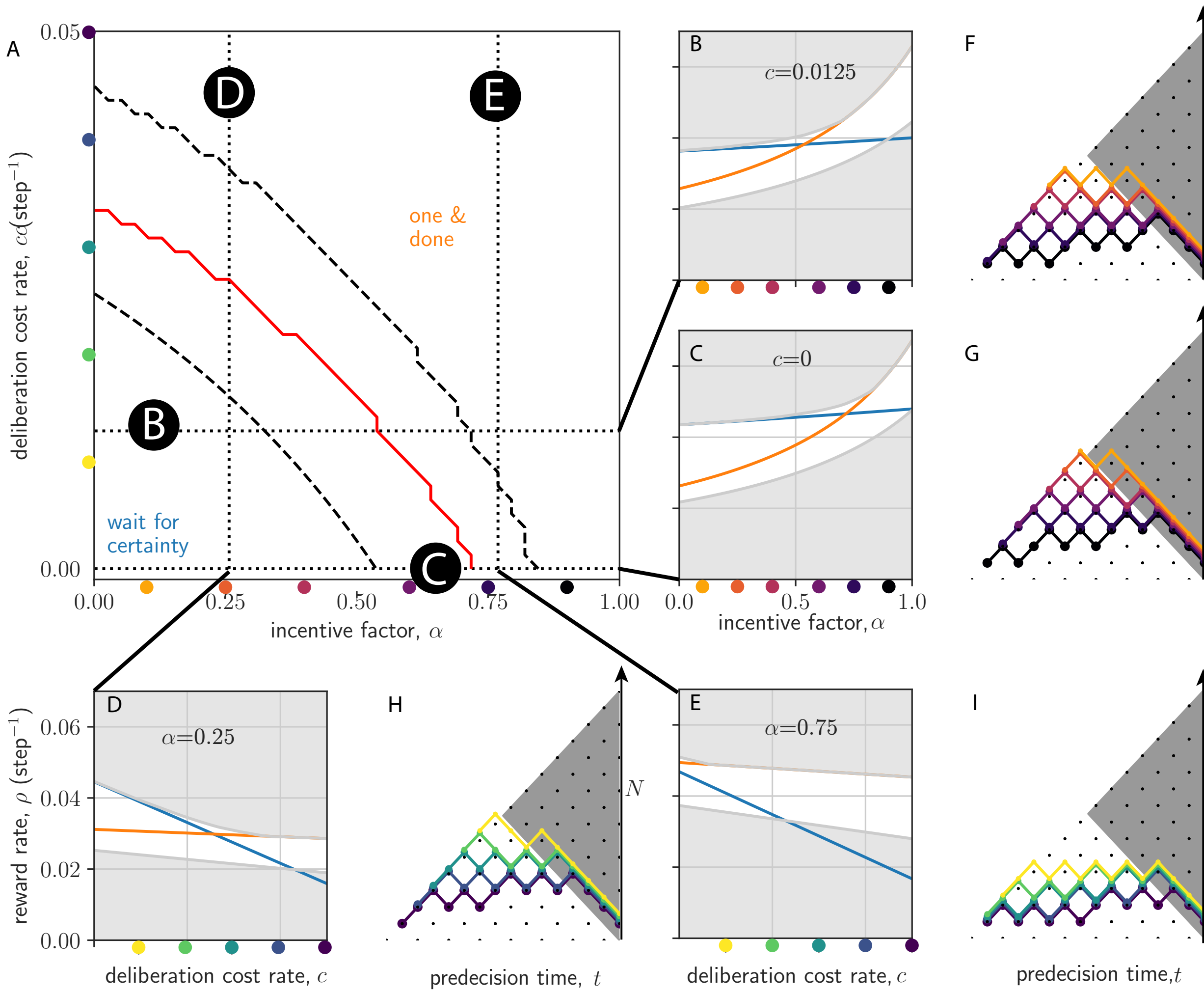



Figure S9. *Reward rate optimal strategies in* $(\alpha, c)$ *plane.* (a) The reward-rate maximizing policy interpolates from the wait-for-certainty strategy at weak incentive (low $\alpha$) and low deliberation cost (low $c$), to the one-and-done strategy at strong incentive (high $\alpha$) and high deliberation cost (high $c$). Dashed lines bound a transition regime between the two extreme strategies. Red line denotes where they have equal performance. (b-e) Slices of the $(\alpha, c)$-plane. Shown are the reward rate as a function of $\alpha$ (b,c) and $c$ (d,e) (wait-for-certainty strategy is shown in blue; one-and-done strategy is shown in orange). $N$ is the magnitude of the token difference

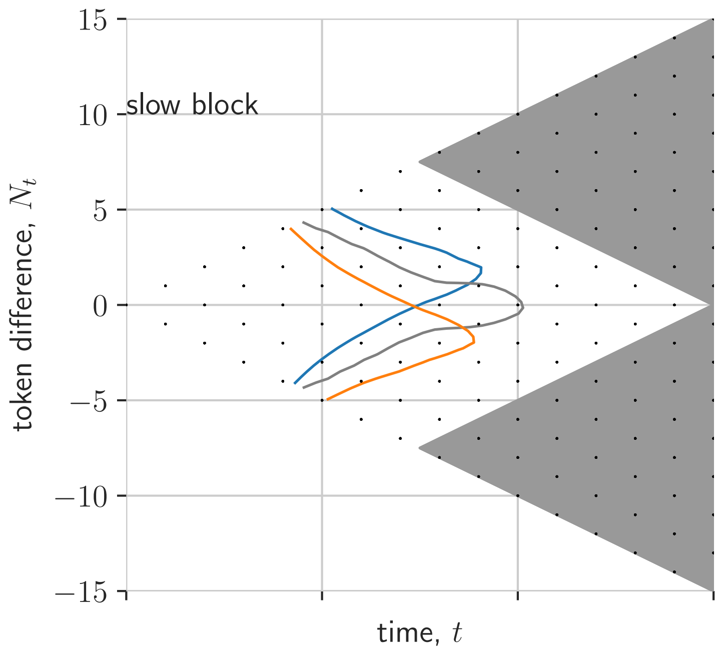



Figure S10. *Asymmetric action rewards skew survival probability.* Here, we plot the half-maximum of the PGD survival probability for three values of the action reward bias, $\gamma = -0.6, 0, 0.6$ (blue, black and orange, respectively). Other model parameters same as in fitted model.

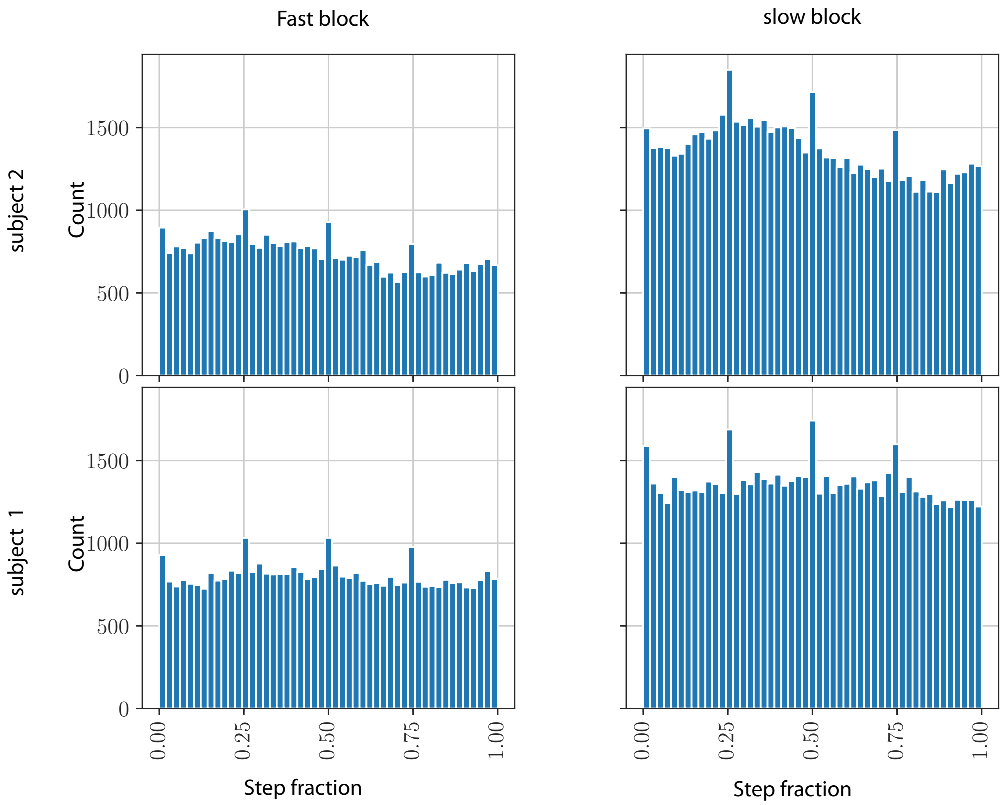



Figure S11. *Decision times relative to token jumps.* Here, we plot the histograms of decision times using their position between token jumps, the step fraction. The data is separated by $\alpha$ and monkey.