Performance-gated deliberation: Urgency as the opportunity cost of time commitment

Maximilian Puelma Touzel
Department of Computer Science and Operations Research, Université de Montréal

Paul Cisek
Department of Neuroscience, Université de Montréal

Guillaume Lajoie
Department of Mathematics and Statistics, Université de Montréal

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 its associated opportunity cost often vary in time and are hard to infer and predict. Here, we propose a non-stationary opportunity cost that inherits the timescales emerging from reward history. It thus readily adapts to changes in context and suggests a generalization of average-reward reinforcement learning (AR-RL) to account for non-stationary contextual factors. We use this opportunity 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. In particular, we propose that opportunity cost is implemented directly as so-called 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 and make readily testable predictions for both neural activity and behaviour.

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

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 kth trial
T_k	duration of kth trial
$t_{\mathrm{dec},k}$	trial decision time
$\mathcal{O}_t$	within-trial opportunity cost
$r_{\max}$	maximum reward acheiveable in a trial
$\langle x \rangle$	trial ensemble average of a quantity x
$\bar{r}(\boldsymbol{S}_t)$	expected reward for reporting at t given $\boldsymbol{S}_t$
$\mathcal{R}_t$	decision regret
ρ	stationary reward rate
ρ^*	optimal stationary reward rate
α	context parameter
ρ_α	context-conditioned stationary reward rate
T_α	context-conditioned stationary average trial duration
$\hat{\rho}_k^\tau$	reward history filtered through a timescale, τ
τ_{long}	a long timescale over which to estimate ρ
τ_{context}	a context-specific timescale over which to estimate ρ_α
ν	tracking cost sensitivity
K	subjective reward scale factor
T_{block}	characteristic duration of a trial block
c	deliberation cost rate
N_t	tokens difference

Table I. Symbol glossary. Highlighted in gray are parameters of the PGD model used 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. It suggests inference over the set of all possible alternatives, the determination of which is ultimately a belief, highly contingent on subjective factors like ambition. There are also alternatives 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 may change over time depending on the context in which a decision is made. Animals will learn to value a given food resource differently depending on whether it is encountered during times of plenty versus time of scarcity, for example. 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 scarce resource to a given use. Nevertheless, the opportunity cost of time is well-studied in relative definitions of value, most notably as the average reward in average-reward reinforcement learning (AR-RL) [2]. AR-RL focusses on deviations from the average reward rather than on discounted reward as in the more widely known discount-reward RL. It was first proposed in neuroscience to extend the reward prediction error hypothesis for phasic dopamine to account also for the observed properties of tonic dopamine levels [3]. AR-RL has since been used to explain human and animal behaviour in foraging [4], free-operant conditioning [5], perceptual decision-making [6, 7], cognitive effort/control [7, 8], and even economic exchange [9]. It is increasingly seen as the more suitable RL formulation for continuing tasks [10]. For one, the average reward is optimized alongside the AR-RL value function. This is in contrast to traditional concepts of fixed accuracy criteria in perceptual decision-making tasks that focus on maximizing reward alone [11]. The decision boundaries emerging as solutions to AR-RL formulations of such tasks collapse in time, limiting deliberation in trials with low return-on-time-investment, e.g. in variable trial difficulty settings [6, 12].

Up to now, however, AR-RL and most of its applications have focussed on fixed context and have used the stationary average reward as the fixed opportunity cost of time, which ignores the above complications. 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. 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 [13]. Consequently, what subjects do within a given trial of the experiment is not just about what happens in that trial, but is also related to the distribution of other trials the subject has seen and can expect to see, which itself could change over the course of a session of the experiment.

Here, we pursue a theory of approximate relative-value decision-making under uncertainty in a setting relevant to decision-making neuroscience. We start by recasting value as a trade-off between opportunity cost and decision regret. Decision regret is the expected reward forfeited by a decision relative to the maximum possible in a trial. Highlighting the risk of value representations in non-stationary environments, we propose an approximation to the AR-RL solution, Performance-Gated Deliberation (PGD), that uses the opportunity cost directly as the collapsing decision boundary, instead of as input to an value optimization problem. Without explicit context knowledge or a value function, PGD trades off speed and accuracy on a given trial according to performance at the longer timescales over which context changes. For subjects that employ PGD, opportunity cost is then directly encoded as "urgency" in the neural dynamics underlying decision-making [6, 14–16]. 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 explicit links to neural mechanisms, we analyze behavior and neural recordings collected over eight years from two non-human primates (NHPs) [17, 18]. 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 [15] and monkeys [18], provides additional support to an existing hypothesis about how neural dynamics implements time-sensitive decision-making [14]. Specifically, neural recordings in monkeys suggest that the evidence needed to make the decision predom-

inates in dorsolateral prefrontal cortex [19]; a growing context-dependent urgency signal is provided by the basal ganglia [20]; and the two are combined to bias and time, respectively, a competition between potential actions that unfolds in dorsal premotor and primary motor cortex [17]. Similar findings have been reported in other tasks - for example, in the frontal eye fields during decisions about eye-movements. As a robust means to balance immediate rewards and the cost of time across multiple timescales, PGD is proposed as the theoretical explanation for why decision-making mechanisms are organized in this way, and as a quantitative model for explaining concurrently recorded behaviour and neural urgency in complex decision-making tasks.

RESULTS

A. Theory of performance-gated deliberation

1. *Opportunity cost, decision regret, and drawbacks of average-reward reinforcement learning*

We consider the general family of tasks consisting of a long sequence of trials indexed by $k = 1, 2, \ldots$ (see fig. 1a). In each trial, a finite sequence, $\boldsymbol{S}_{t_{\max}} = (S_0, \ldots, S_{t_{\max}})$, of states, S_t, $t = 0, \ldots, t_{\max}$, 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 report, $t_{\mathrm{dec},k}$, and the chosen option determine both the reward received, R_k, and the trial duration, $T_k \geq t_{\mathrm{dec},k}$. 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. For a fixed strategy, the *stationary reward rate* (see slope of dashed line in fig. 1a(right)) is

$$\rho := \lim_{k\to\infty} \sum_k R_k \Big/ \sum_k T_k \text{ (time-average)}. \tag{1}$$

For a stochastic environment, this definition of ρ includes an ensemble average over that randomness. Free-operant conditioning, "patch leaving", and several perceptual decision-making tasks often fall into this class. Previous work [6, 21] has studied the choice probability $P(R = r|\boldsymbol{S}_t, t_{\mathrm{dec}} = t)$, for which the belief of correct report for binary rewards, $P(R = 1|\boldsymbol{S}_t, t_{\mathrm{dec}} = t)$, equals the expected trial reward, denoted $\bar{r}(\boldsymbol{S}_t)$ [6] (see [22] for more about the relationship between value-based and perceptual decisions). We consider greedy strategies that report the choice with the largest belief at decision time. The decision problem is then about when to decide.

Average-reward reinforcement learning (AR-RL), first proposed in artificial intelligence [23], was later incorporated into reward prediction error theories of dopamine signalling [3] and employed to account for the the opportunity cost of time [5]. AR-RL was subsequently used to study reward-based decision-making in neuroscience and psychology [6, 7, 24, 25]. Through maximization (in expectation) of the average-adjusted future return, $\sum_{t'>t}(R_{t'} - \rho)$, where ρ (eq. (1)) is either estimated online or obtained self-consistently, AR-RL algorithms aim to achieve the highest ρ possible. In an trial-based task, the focus of AR-RL narrows onto the return in a single trial [6]. To connect to opportunity cost, we now provide an equivalent perspective in terms of costs to the agent.

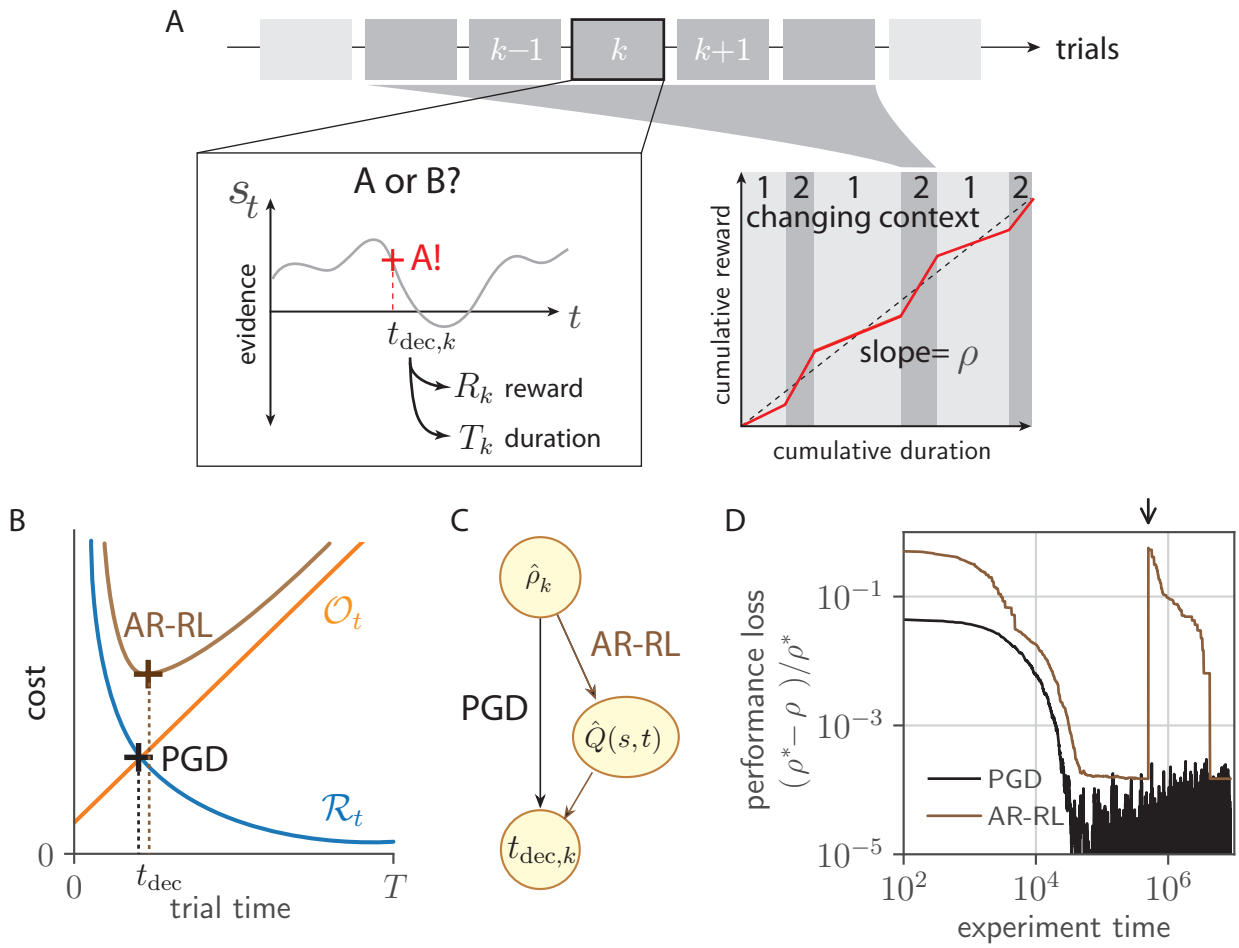

Figure 1. *AR-RL and Performance-Gated Deliberation.* (a) Task setting. Left: Within trial evidence, S_t evolves over trial time t in successive trials indexed by k. The decision 'A' is reported at the decision time $t_{\mathrm{dec},k}$ (red cross), determining reward, R_k, and trial duration, T_k. Right: Cumulative reward versus cumulative duration. Reward rate (slope of red line), varies with alternating context (labelled 1 and 2) around average reward, ρ (dashed line). (b) Decision rules based on regret, $\mathcal{R}_t$ and opportunity cost, $\mathcal{O}_t$. The AR-RL rule (brown cross) finds t that minimizes $\mathcal{O}_t+\mathcal{R}_t$. The PGD rule (black cross) finds t at which they intersect, $\mathcal{O}_t = \mathcal{R}_t$. (c) Schematic diagram of each algorithm's dependency. PGD computes a decision time directly from the estimated reward rate only, while the AR-RL first estimates a value function, whose optimization produces the decision time. (d) Loss over learning time in a patch-leaving task. Loss is relative error in performance with respect to the optimal policy, $(\rho^* - \rho)/\rho^*$ (AR-RL: brown, PGD: black). The arrow indicates when the state labels were randomly permuted.

We define the *decision regret* at time t within a trial as the difference,

$$\mathcal{R}_t = r_{\max} - \bar{r}(\boldsymbol{S}_t)\ , \tag{2}$$

where $r_{\max}$ is the maximum trial reward possible *a priori*. An agent lowers its regret towards zero by accumulating more evidence, i.e. by waiting. Waiting, however, incurs opportunity cost: the reward lost by not acting. We denote $\mathcal{O}_t$ as the opportunity cost incurred up to a

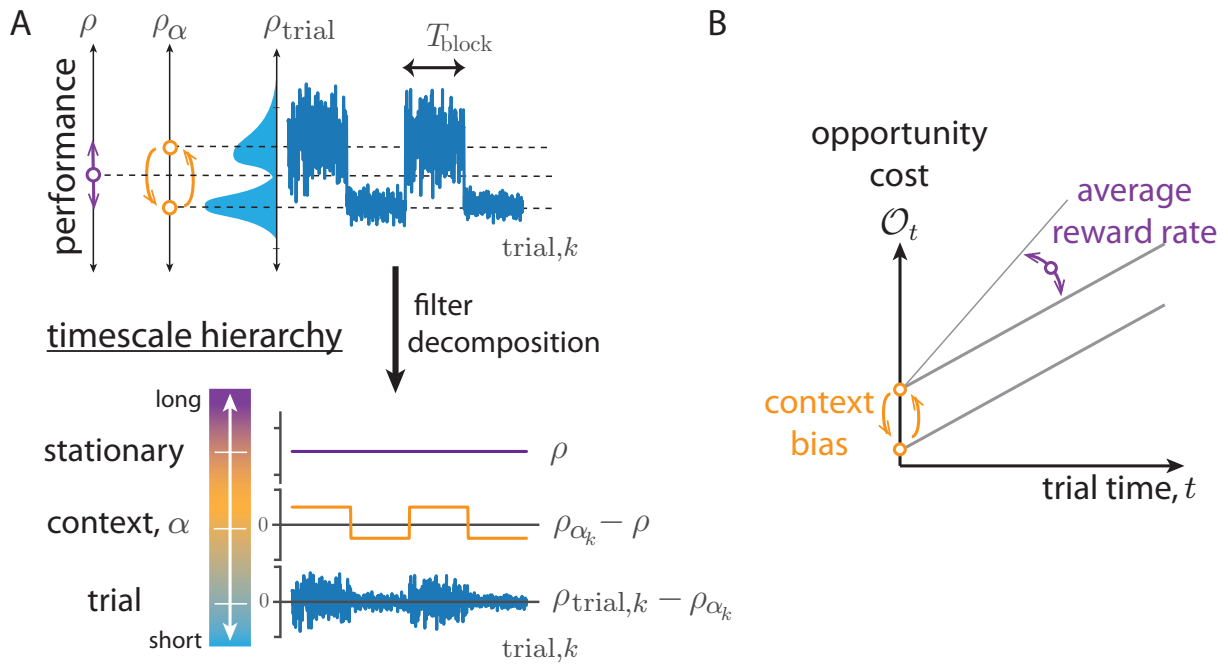

Figure 2. *Non-stationary opportunity cost.* (a) Variation of trial performance ($\rho_{\text{trial},k} := R_k/T_k$), context performance ($\rho_\alpha = \langle \rho_{\text{trial},k} \rangle_{k|\alpha}$), and effectively stationary performance ($\rho \sim \langle \rho_{\text{trial},k} \rangle_k$) in an experiment are decomposed into a hierarchy by filtering reward history on trial, context, and long timescales, respectively. (b) The corresponding trial opportunity cost from (d) grows with slope ρ, and is offset by the context-aware cost deviation (see eq. (5)).

time t in a trial. For $T_k = t_{\text{dec},k}$, the opportunity cost of time in AR-RL is

$$\mathcal{O}_t = \rho t \ . \tag{3}$$

With these definitions, the average-adjusted trial return for deciding at a time t can be expressed as $r_{\max} - (\mathcal{R}_t + \mathcal{O}_t)$, such that it is maximized by jointly minimizing $\mathcal{O}_t$ and $\mathcal{R}_t$ (brown cross in fig. 1b), giving the AR-RL optimal solution (see Methods for a formal statement and solution of the AR-RL). More generally, this perspective emphasizes that an agent's solution to the speed-accuracy trade-off is about how it balances decaying regret and growing opportunity cost.

Using value representations in decision-making can be a liability in real world tasks where task statistics can vary unpredictably. To illustrate this, we consider the following foraging task. An animal 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. Indexing patches with s, the food return is analogous to $\bar{r}(s,t)$ (c.f. eq. (2)), but directly observed and deterministic given s, which here serves as context. To perform well, the animal needs to decide when to move on from depleting the current patch (see Methods for further details about the task and its solution). 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 approaches that of the optimal policy in time (brown line) as the estimation of the AR-RL trial return improves with experience. However, if the environment undergoes a significant perturbation (e.g. a random permutation of the state labels at the time indicated by the arrow in fig. 1d), the performance of this AR-RL algorithm drops approximately back to where it started. More generally, any approach that learns by relying on explicit state associations shares this drawback, which includes those approaches that directly learn policies instead of value functions [26]. Could high-value decision times be obtained without having to associate value or action to state?

2. *Performance-Gated Deliberation*

We propose that instead of maximizing average reward as in AR-RL (equivalent to minimizing $\mathcal{O}_t + \mathcal{R}_t$), the agent simply takes as its decision criterion the intersection of opportunity cost $\mathcal{O}_t$ and decision regret $\mathcal{R}_t$ (shown as the black cross in fig. 1b).

$$t_{\text{dec}} := \min_t \{t \mid \mathcal{O}_t \geq \mathcal{R}_t\} \qquad \text{(PGD decision rule)} \tag{4}$$

We call this heuristic rule at the center of our results *Performance-Gated Deliberation* (PGD). Plotted alongside the AR-RL performance in fig. 1d, PGD (black line) achieves better performance than AR-RL overall in our example foraging task and is insensitive to the applied disturbance since it uses $\mathcal{O}_t$ and $\mathcal{R}_t$ directly when deciding rather than as input to a value optimization (fig. 1c).

We constructed the above task such 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, the animal will have to learn the state-reward associations in the expected reward, $\bar{r}(\boldsymbol{S}_t)$, over the residual uncertainty in the trial. This lower-level learning is more likely to be stationary across trials, independent of context, and we study such a case here.

3. *Temporal reward filtering for a dynamic opportunity cost of time*

The state perturbation 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 opportunity cost is thus needed–one that can account for extended timescales over which performance varies beyond the moment-to-moment. To address this, we leverage the fact that an effective decision hierarchy naturally emerges from tasks with multiple timescales. Performance variation over this hierarchy then serves as a means to construct a non-stationary opportunity cost of time. We illustrate an example using the task we will present in detail in the following section. This task has a context parameter, α, whose variation in time adds a timescale beyond the moment-to-moment and can serve as a source of non-stationarity.

In this example, 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, ρ (purple), with context variation due to the switching, as well as context-conditioned trial-to-trial variation. The time-average over recent performance can be estimated by filtering the sequence of rewards, e.g. through a low-pass filter, with an

integration time, τ_{context}, tuned to trade-off the bias and variance of the estimate. More generally, filtering performance history through a hierarchy of timescales decomposes this performance variation into a hierarchy of timescale-specific components (fig. 2b (bottom)). Any performance variation contributed by a given component can be interpreted as the result of planning on that timescale, intended or not. Opportunity costs of time are now specific to the level in the decision-making hierarchy and, crucially, are incurred on their respective timescale. They nevertheless combine into a single dynamic opportunity cost consistent with this decision hierarchy. How can these filtered components be obtained in practise?

The stationary reward rate, ρ, can be estimated by the agent to high precision from applying a low pass filter with a long integration time, τ_{long}, to the reward sequence R_k [7, 27]. We denote this estimate $\hat{\rho}_k^{\tau_{\text{long}}}$ (See Methods for estimator details). If α_k were a constant sequence, $\mathcal{O}_t = \hat{\rho}_k^{\tau_{\text{long}}} t$ and we recover eq. (3) of AR-RL. Context varies, however, on a specific timescale and estimating performance on this timescale then leads to a second filtered estimate $\hat{\rho}_k^{\tau_{\text{context}}}$ that, unlike $\hat{\rho}_k^{\tau_{\text{long}}}$, tracks the effective instantaneous, context-specific performance, ρ_{α_k}. The estimation error associated with the estimate $\hat{\rho}_k^{\tau_{\text{context}}}$ arises from a trade-off, controlled by the value of τ_{context}, between its bounded speed of adaptation and its finite memory.

Using $\mathcal{O}_t = \hat{\rho}_k^{\tau_{\text{context}}} t$ (c.f. eq. (3)) at first appears like a straightforward way to extend the AR-RL formulation of stationary opportunity costs to the non-stationary case. However, this incorrectly lumps together two distinct types of opportunity costs: those incurred moment-by-moment and those incurred as a result on planning over context variation. Instead, the opportunity cost consistent with the implicit decision hierarchy (c.f. fig. 2a) is

$$\mathcal{O}_t = \rho t + (\rho_\alpha - \rho) T_\alpha \ . \quad \text{(context-aware opportunity cost)} \tag{5}$$

Equation (5) is plotted over trial time t in fig. 2b. Its first term is the conventional AR-RL contribution from the moment-to-moment opportunity cost of actions using the stationary reward rate, ρ. The second, novel term in eq. (5) is a baseline cost incurred at the beginning of each trial and computed as the average deviation in opportunity cost accumulated over a trial from that context (T_α is the average duration of a trial in context α). This deviation fills the cost gap made by using the stationary reward rate ρ in the moment-to-moment opportunity cost instead of the context-specific average reward, ρ_α. This baseline has 0-mean, as verified through the mixed context ensemble average reward (e.g. $\rho \equiv \sum_\alpha \rho_\alpha T_\alpha / \sum_\alpha T_\alpha$ when trials from different context are uniformly sampled such that $\sum_\alpha (\rho_\alpha - \rho) T_\alpha = 0$). We estimate this baseline cost using $(\hat{\rho}_{k-1}^{\tau_{\text{context}}} - \hat{\rho}_{k-1}^{\tau_{\text{long}}}) T_{k-1}$, where we have used the sample T_{k-1} in lieu of the average T_α. See fig. S1 for a signal filtering diagram that estimates eq. (5) from reward history.

B. Neuroscience application: PGD in the tokens task

In this section, we apply the PGD algorithm to the "tokens task" [15]. We first give an illustrative example with periodic context dynamics, and then an application to a set of non-human primate experiments in which context variation was non-stationary [18]. For the latter, we used the relaxation timescales after context switches of the decision times to fit the model. We then validate the model on the remaining features of the data. In particular, we assessed PGD's ability to explain the concurrently recorded behaviour and neural activity in

premotor cortex (PMd) of the two monkeys via their context-specific behavioural strategies and the temporal profile of their underlying neural urgency signals, respectively.

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 of 15 tokens. With the t^{th} jump labelled $S_t \in \{-1, 1\}$ serving as the state, we process the history of states 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). Importantly, after the subject reports, the interval between jumps contracts to once every 50ms (the "fast" condition) or once every 150ms (the "slow" condition), giving the subject the possibility to save time by taking an early guess. The interval contraction factor, $1-\alpha$, for the fast ($\alpha = 3/4$) and slow ($\alpha = 1/4$) condition is parametrized by $\alpha \in [0, 1]$, 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. The tokens difference, N_t, provides sufficient evidence to compute the belief (equivalent to the expected reward) that one or the other target will have more tokens by the end. From this, the subject can compute the expected decision regret (derived in Methods). We display this regret dynamics in fig. 3b. It evolves on a lattice (gray), always starting at 0.5 and ending at 0. We assume the agent has learned to track this decision regret. The PGD agent uses this regret, along with the estimate of the context-aware opportunity 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 α switches back and forth every 300 trials (see fig. 3c). We call such segments of contiguous context 'trial blocks', with context then alternating between slow ($\alpha = 1/4$) and fast ($\alpha = 3/4$) blocks.

The decision boundary in PGD is implemented in cost space and given by the opportunity cost(fig. 3b). We later specify the equivalence with the previously studied implementation in belief space [6]. Opportunity cost is estimated using performance estimates on context-specific and long timescales, such that it depends on performance history and is therefore dynamic. In order for the two performance filters designed for this purpose to generate reasonable estimates, we set τ_{context} to average over many (10) trials and τ_{long} two orders of magnitudes larger so that the long timescale filter averages over many blocks. In the case of periodic context, the performance estimates 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 (fig. S2d). The two estimates behave differently from one another because of their distinct integration times. For example, the context estimate $\hat{\rho}_k^{\tau_{\text{context}}}$ relaxes relatively quickly after context switches to the context-conditioned stationary average performance, but exhibits stronger fluctuations as a result (dashed lines in fig. 3d; c.f. fig. S2e). 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. In general, when the block duration, T_{block}, is much less than τ_{long} ($T_{\text{block}}/\tau_{\text{long}} \ll 1$), the within-block exponential relaxation is roughly linear and so the average unsigned deviation between $\hat{\rho}_k^{\tau_{\text{long}}}$ and the actual stationary reward, ρ, is $1 - \exp\left[-T_{\text{block}}/\tau_{\text{long}}\right] \approx T_{\text{block}}/\tau_{\text{long}} \ll 1$. This scaling fits the simulated data well (fig. S2d:

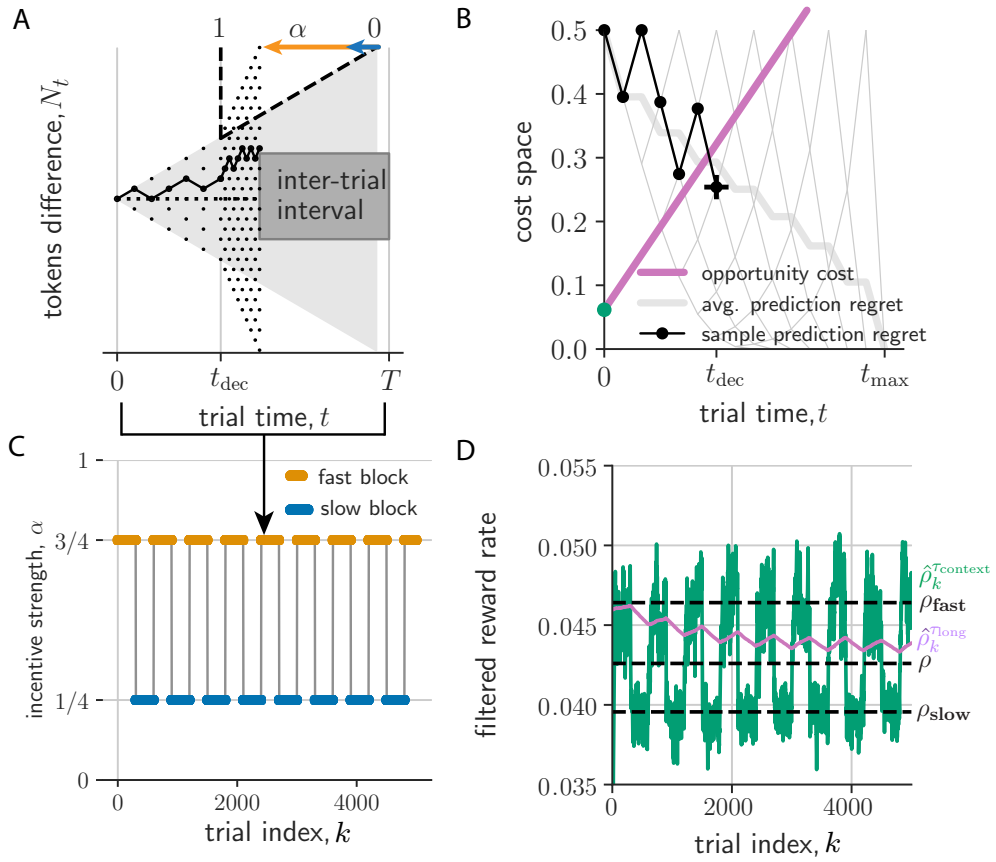

Figure 3. *PGD agent performs the tokens task for periodic context switching.* (a) A tokens task trial for $\alpha = 3/4$ and decision time t_{dec}. Plotted is tokens difference, N_t, vs. trial time, t. The trial duration is T, which includes an inter-trial interval. (b) Decision dynamics in cost space obtained from evidence dynamics in (a). Expected decision regret trajectories (gray lattice; thick gray: trial-averaged) start at 0.5 and end at 0. The one from (a) is shown in black. t_{dec} is determined by the crossing of the regret and opportunity cost (purple). (c) Incentive strength switches between two values every 300 trials. (d) Expected rewards filtered on τ_{long} ($\hat{\rho}_k^{\tau_{\text{long}}}$, purple) and τ_{context} ($\hat{\rho}_k^{\tau_{\text{context}}}$, green). Black dashed lines from bottom to top are $\rho_{\alpha=1/4}$, ρ, and $\rho_{\alpha=3/4}$.

inset).

The dynamics of these two performance estimates drives the dynamics of the decision times via the expression of the decision boundary (eq. (5)). The decision time sequence relaxes after a context switch (fig. S2c) to the context-conditional average but exhibit strong fluctuations due to the sequence of random walk realizations. The result of these dynamics is that the PGD algorithm sacrifices accuracy to achieve shorter trial duration in trials of the the fast block, achieving a higher context-conditioned reward rate compared to decisions in the slow block (the slopes shown in the inset of fig. S2d).

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

Next, we apply to the PGD algorithm to the actual context-switching α-sequence used in the experiments reported in [18] and fit the model to the recorded behaviour of the two

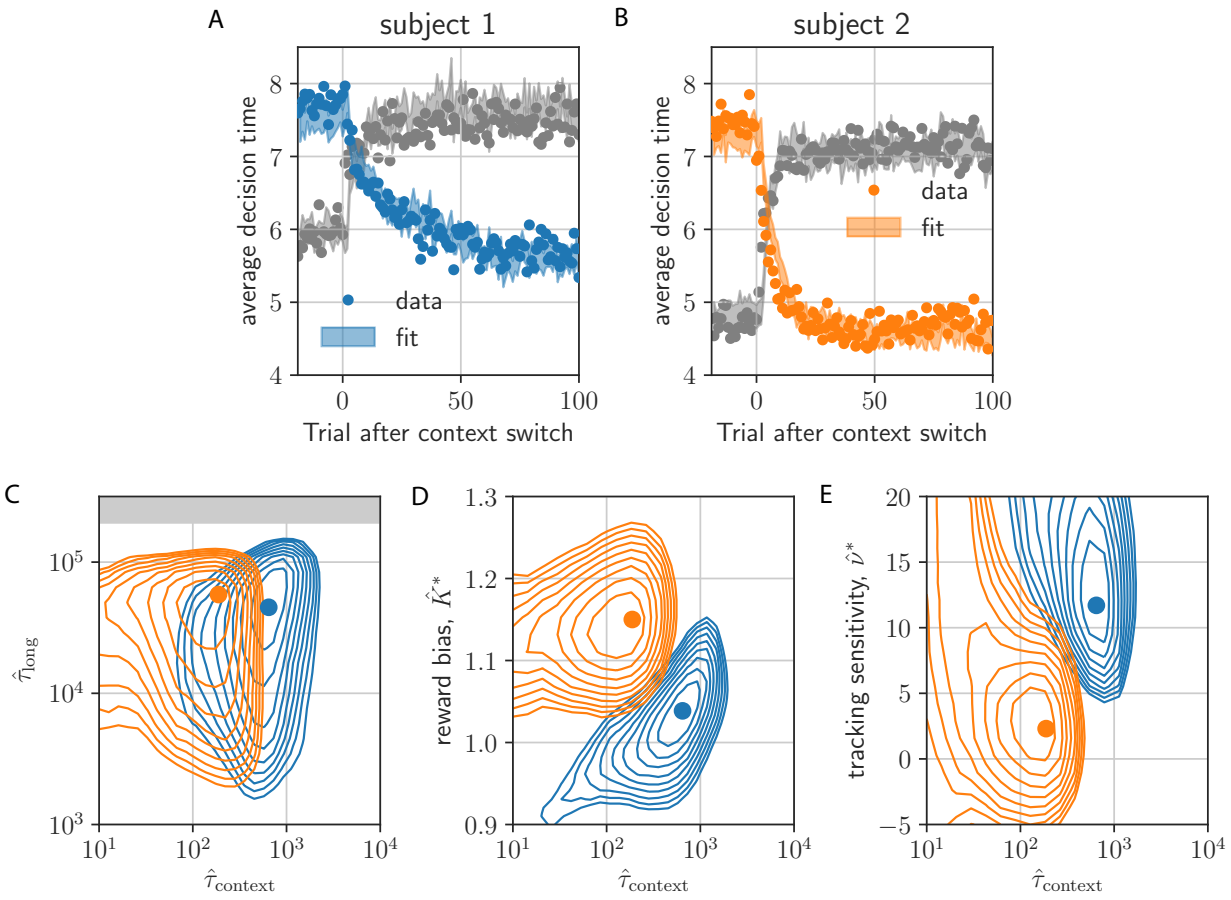

Figure 4. *Model fit.* (a,b) decision times (dots) aligned on the context-switching event and averaged. Shaded region are standard error bounds of model fit to each of the two types of switching events. (c) $(\hat{\tau}_{\text{context}}, \hat{\tau}_{\text{long}})$-plane cut through error hypersurface at $\nu = \hat{\nu}^*$ and $K = \hat{K}^*$ (gray area indicates timescales within an order of magnitude of the duration of the experiment). Contours show the first 10 contours incrementing by 0.01 error from the minimum (circle markers). Colors as in (a) and (b). (d) Same for $(\tau_{\text{context}}, \hat{K})$ at $\hat{\tau}_{\text{long}} = \hat{\tau}^*_{\text{long}}$ and $\nu = \hat{\nu}^*$. (e) Same for $(\tau_{\text{context}}, \hat{\nu})$ at $\hat{\tau}_{\text{long}} = \hat{\tau}^*_{\text{long}}$ and $K = \hat{K}^*$.

monkeys. As with the above example (c.f. fig. 3), trials were structured in alternating blocks of two values of α, but here the blocks exhibit large, irregular fluctuations in size, primarily from the experimenter adapting the protocol according to fluctuations in motivation of the subject (see fig. 5a)[28].

So far, PGD has only two free parameters: the two filtering time constants, τ_{long} and τ_{context}. We anticipated only a weak dependence of the fit on the τ_{long}, so long as it exceeded the average duration of a handful of trial blocks enabling a sufficiently precise estimate of ρ. In contrast, the context filtering timescale, τ_{context}, is a crucial parameter as it dictates where the PGD agent lies on a bias-variance trade-off in estimating ρ_{α_k}, the value of which determines the context-specific contribution to the opportunity cost. With the addition of a subjective reward bias factor, K, and a tracking-cost sensitivity parameter, ν, we could quantitatively match the average decision time dynamics around the two context switches (fig. 4a,b; see Methods for fitting details). Inspecting the error function surface around

the fitted parameter values between the two monkeys suggests a constrained relationship between K, τ_{context}, and ν (fig. 4c-e). In particular, the faster the adaptation, the larger the reward bias and the lower the sensitivity to the tracking cost. We also note that, as expected, the data precisely identified τ_{context}, but only set a soft lower bound on the value of τ_{long} (fig. 4c). The resulting temporal statistics of the behaviour for these fitted parameters (e.g. temporal correlations) gave good correspondence with the data (see fig. S6).

With the model fit to average decision times around switching events, we could then test the model on the state-dependence of the 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 details). We give this conditional probability for each context for subject 1 in fig. 5b-e. We left behavioural noise sources out of the model in order to more clearly demonstrate the behaviour of PGD algorithm. As a result, the model underestimates the spread of probability over time and tokens state. Nevertheless, the remarkably smooth average strategy is well captured by the model (white dashed lines in fig. 5c,e). Specifically, fast block strategies (fig. 5d,e) are shifted to earlier decision times by similar amounts relative to slow block strategies (fig. 5d,e) in both model and data. Subject 2 was qualitatively similar, but shifted slightly to earlier times (fig. S3). Our model explains this difference as resulting from subject 2's larger reward bias and faster context integration (c.f. fig. 4d). The correspondence between model and data in this high-dimensional space of behaviours across subjects is remarkable given that the model has essentially only a single, subject-specific degree of freedom (τ_{context}), which we have fit using a timescale that is independent of the state dynamics.

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 context, α (here $\alpha \in [0, 1]$). We added to the reward objective a constant deliberation cost rate, c, incurred during the deliberation period in each trial [6]. A movement cost, i.e. a constant cost incurred by either of the reporting actions, has a qualitatively similar effect (data not shown). Solving a value iteration problem via dynamic programming 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 α is shown in fig. 5f for $c = 0$. The optimal solutions generating these reward rates interpolate from the wait-for-certainty strategy at low α to the one-and-done strategy [29] at high α (this holds also for $c > 0$ and thus over the entire (α, c)-plane; see fig. S8 for the complete dependence). The performance of the α-conditioned reward rates achieved by the two primates and a reference human are also shown in fig. 5f. As expected, they fall between the optimal strategy and the strategy that picks one of the three actions (report left, report right, and wait) at random. Given the good match in behaviour between model and data (c.f. fig. 5b,e), this intermediate performance is shared by the stationary PGD agent. Moreover, the fitted PGD model is able to capture the primate behaviour by resolving the residual ambiguity ($N_t \approx 0$) at intermediate trial times (fig. 5b-e). In contrast, all of the optimal strategies across c give no intermediate decision times at ambiguous ($N_t \approx 0$) states, invariably waiting until the ambiguity resolves. Thus, whereas optimal policies shift around the edges of the relevant decision space as α 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 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 deliberation cost

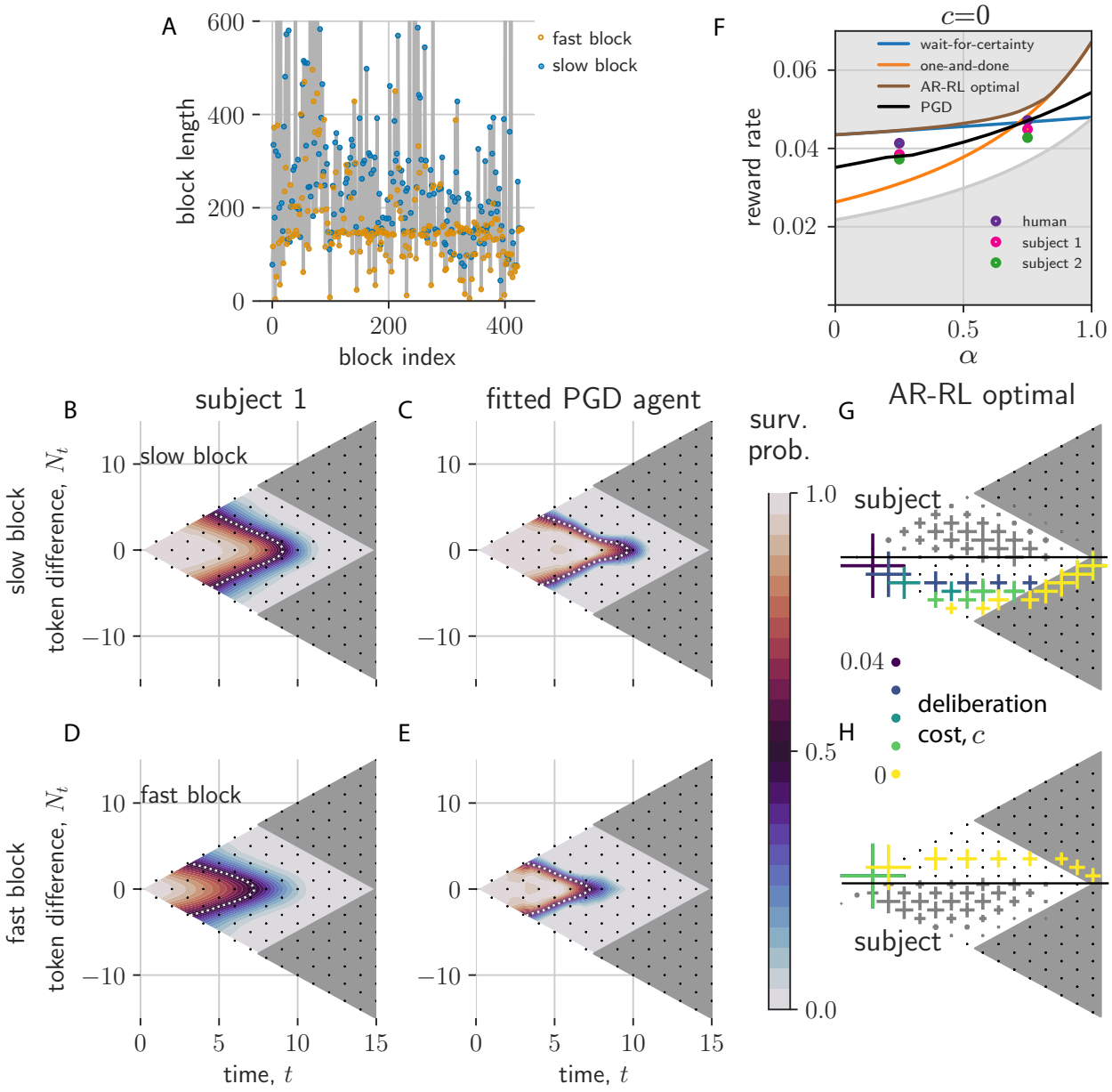

Figure 5. *Comparison of PGD and NHP data in non-stationary α dynamics from Ref. [18].* (a) Block length sequence used in the experiment (c.f. fig. 3c). (b-e) Interpolated state-conditioned survival probabilities, $P(t_{\text{dec}} = t|N_t, t)$, over slow (b,c) and fast (d,e) blocks. White dashed lines 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, α, and no deliberation cost ($c = 0$) (wait-for-certainty strategy shown in blue; one-and-done strategy shown in orange). We additionally show the context-conditioned reward rates for the two primates as well as a reference human, and the PGD algorithm (black line). Reward rates for primates are squarely in between the best and uniformily random strategy (lines bounding the upper and lower gray regions, respectively). (g,h) Decision time probability histograms from optimal decision boundaries across different values (colored crosses) of the deliberation cost 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 decision time frequencies for comparison. Cross size corresponds to histogram frequency.

rate.

3. *Neural urgency and opportunity cost*

So far, we have focussed on fitting and analyzing the PGD model with respect to the recorded behaviour. Here, we take a step in the important direction of confronting the above theory of behaviour with the measured neural dynamics that we propose drive it. The proposal mentioned at the end of the introduction has evidence strength and urgency combining in PMd, whose neural dynamics implements the decision process in the tokens task. 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); [6, 30]), the rising choice probability collides with the collapsing boundary. In the equivalent regret and opportunity cost formulation developed here (fig. 6a(middle)), the falling regret collides with the rising opportunity cost. The collapsing boundary in belief space can be parametrized as $C - u_t$, where C is the initial strength of belief, i.e. 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 denotes the choice probability $p(R = 1|\boldsymbol{S}_t, t_{\text{dec}} = t)$. For AR-RL optimal policies, u_t has a complicated dependence on the opportunity cost sequence, $\mathcal{O}_t$. For PGD, in contrast, C is interpreted as the maximum reward r_{max} and u_t is identically $\mathcal{O}_t$. 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 recorded data of [18] in [31]. 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), $\tilde{u}_t$ is a signal arising from a population-level drive to all PMd neurons, irrespective of their selectivity. The urgency, $\tilde{u}_t$, can then be extracted from neural recordings by conditioning on zero-evidence states ($\tilde{b}_t = 0$) and averaging over cells. In fig. 6b, we replot their result (c.f. fig.8b of [31], which averages over the two monkeys) and overlay the mean (+/- standard deviation) of the opportunity cost sequence, $\mathcal{O}_t$ (shaded area in fig. 4; averaged over all trials of the two fitted PGD models for each context). To facilitate our qualitative comparison, we convert reward to spikes/step by simply adjusting the y-axis of the opportunity cost. The observed urgency signals then lie within the uncertainty of the context-conditioned opportunity 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; (3) the baseline offset between conditions, where the fast condition is offset to higher values than the slow condition. In the absence of a theory, such features remained descriptive. Now, each has a specific meaning by interpreting urgency as opportunity cost: (1) the animal uses a constant opportunity 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) context-aware planning leads to an opportunity cost baseline offset with sign given by the deviation $\rho_\alpha - \rho$ from the stationary average, ρ. Taken together, the data support the interpretation that neural activity driving context-conditioned behavioural responses is gated by opportunity cost, with earlier responses in high reward rate contexts because the

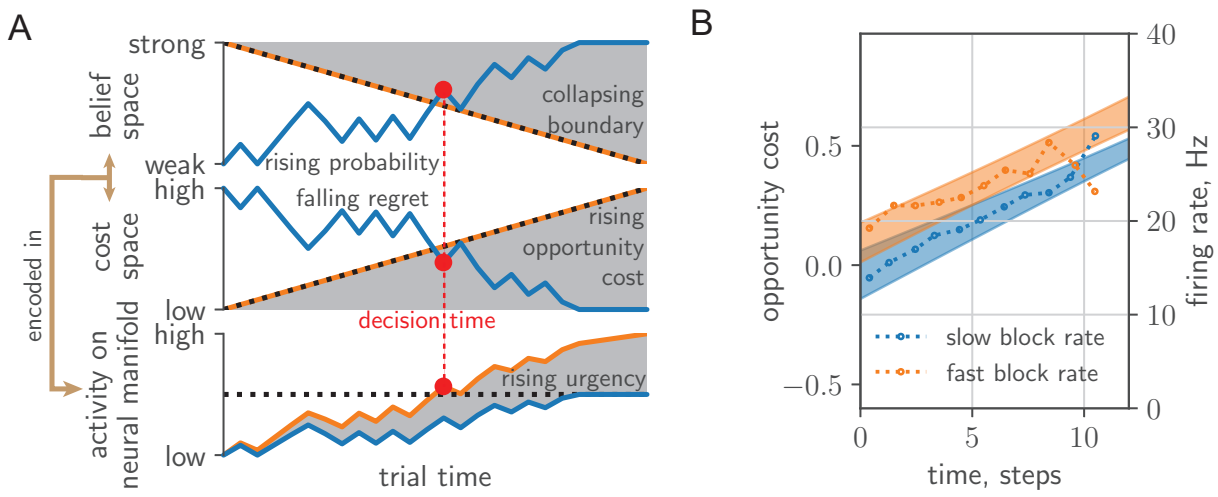

Figure 6. *Neural urgency and collapsing decision boundaries and comparison with data.* (a) Top: Rising probability of success (black) meets collapsing decision boundary (green) in belief space. Middle: Falling regret (black) meets rising opportunity cost (green) in cost space. Bottom: Belief/regret is encoded (black) into a low-dimensional neural manifold, with the addition of an urgency signal (green). (c.f. fig.8 in [6].) (b) Opportunity cost linearly maps onto the urgency signal extracted from zero-evidence conditioned cell-averaged firing rate in PMd (c.f. fig.8b from [31]).

opportunity cost of deliberation there is higher.

This interpretation was made possible because of the prescriptive nature of PGD. In contrast, the nature of urgency has remained largely unexplored in normative approaches, which are limited by the absence of an interpretable form for urgency. Moreover, our analyses using the PGD algorithm provided us with a trial-by-trial account of the behaviour. We used the temporally structured data to fit PGD and explain the observed transition dynamics across context switches. Existing normative approaches like AR-RL say little about such phenomena because they focus on a single trial realization.

DISCUSSION

We have proposed PGD, a heuristic decision-making algorithm that gates deliberation based on performance. We constructed a foraging example for which 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. PGD is widely applicable and at once 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 and tracking one's own performance in that environment. This decomposition naturally maps onto how we believe brains make decisions: a cortico-basal ganglia system in which performance estimated on multiple, behaviourally-relevant timescales is broadcast to multiple decision-making areas to gate the speed of their respective attractor-based decision-making dynamics. 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. We used behavioural data to shape the theory, and neural recordings to provide evidence of one of the neural correlates it proposes: the temporal profile of neural urgency. 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.

Scientific and clinical implications While the view that tonic dopamine encodes average reward is two decades old [3], the multiple timescale representation has received increasing empirical support in recent years, from cognitive results [32–34] to a recent unified view of how dopamine encodes reward prediction errors using multiple discount factors [35] and of dopamine as encoding both value and uncertainty [36]. Dopamine's effect on time perception has been proposed [37] and experimentally supported [38], but its putative effect on decision speed, the mechanism by which speed is implicated in the neural dynamics of decision-making areas driving motor responses was unknown. Our theory fills this explanatory gap by considering dynamic evidence tasks and mechanistically identifying urgency as the means by which the neural representation of reward ultimately affects neural dynamics in decision-making areas. It suggests specific driving signals, for example, to be incorporated in attractor-based decision-making models [30, 39, 40]. PGD is also an example of decision-making without an explicit value function, contributing to a contemporary debate that questions their necessity in the neuroscience of decision-making [26]. Psychological test batteries are demonstrating the role of urgency as a transdiagnostic indicator of reward and motor processing impairments. 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 [41–43]. We provide a concrete proposal for a signal filtering system that extracts a context-sensitive opportunity cost from a temporal difference error sequence putatively encoded by dopamine. The exact neural substrate for this system remains to be identified.

Regret estimation Beyond the estimation of average opportunity cost (see [44] for an example that also tracks performance fluctuations), we assumed that the agent had a precise estimate of the expected reward. This was used to compute the within-trial decision regret. For the tokens task, a recorded signal in dorsal lateral prefrontal cortex of non-human primates correlates strongly with success probability [19], which in this particular case is equivalent to the expected reward. 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 [36]. 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 [45]. The expected reward at a chosen future time can be easily decoded from this representation using a linear classifier.

Theory predictions A distinguishing feature of our decision-making theory is its prescriptive character, e.g. with regards to behaviour via the shape of the action policy (c.f. fig. 5b-h). PGD varies markedly with reward structure and thus provides a wealth of predictions for how observed behaviour should be altered by reward structure. 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 in task structure for which the theory predicts a distinctly asymmetric shape using our survival probability representation (fig. S9; for details see Methods). Our theory is also prescriptive for neural activity via the temporal profile of neural

urgency. The data we analyzed was for block lengths short enough that the slope of $\mathcal{O}_t$ remains fixed across blocks. On the other hand, if $T_{\text{block}}/\tau_{\text{long}} \gg 1$, $\rho_k^{\tau_{\text{long}}}$ approaches ρ_α except when undergoing large, transient excursions after context switches. Thus, the opportunity 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 should exhibit increasing variation across block type with longer blocks.

Reinforcement learning theory Our work impacts reinforcement learning theory by describing how to generalize average-adjusted value functions to non-stationary opportunity costs. The epistemic perspective entailed in the estimation of these costs parallels the recent interpretation of the discount-reward formulation as encoding knowledge about the volatility of the environment [46]. Our work also suggests a new class of RL algorithms between model-based and model-free, in that only parts of the algorithm need adjustment upon task structure variation, reminiscent of how the effects of complex state dynamics are decoupled from reward when using a successor representation [47]. Being based in the average-reward rather than the discount-reward formulation of RL, however, this generalization is perhaps better suited to the continuing task setting of clear relevance to many biological agents. We have left a detailed algorithmic analysis of this 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 [48], most do not. The structures underlying the hard-coded neural implementation we offered for PGD is certainly shared with humans, but the degree to which we exploit PGD requires further study. Hard-coded or not, the question remains if PGD is optimal with respect to some bounded rational objective. In spite of the many issues with the latter approach [49], using it to further understand the computational advantages of PGD is an interesting direction for future work.

Humans, despite our apparent access to sophisticated computation, still exhibit measurable bias in how we incorporate past experience [50]. 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 [51]. In that work, numerical experiments demonstrated that this strategy outperforms at a population level the optimal Nash equilibrium for this game, demonstrating that such seemingly sub-optimal strategies can have surprisingly good evolutionary benefit. 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 individual and evolutionary timescales have shaped decision-making in non-stationary environments.

METHODS

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 $\mathcal{O}_t > r_{\max} - \mathcal{R}_t = \bar{r}(s,t)$ [4]). 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 setting 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 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 their convergence times.

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 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 λ, 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 hereon 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, $r - \rho t$, at the center of AR-RL. Given the return profile we have chosen, the corresponding optimal decision time, t_s^*, in the sth 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) costs more. We chose this return profile such that stationary PGD learning gives exactly the same decision times, i.e. the condition $\mathcal{O}_t = \mathcal{R}_t$ for patch s here takes the form $\rho t_s = r_{\max,s}/(\lambda t_s)$. Thus, they share the same optimal reward rate, ρ^*. Using t_s^* for each patch in eq. (6) gives a self-consistency equation for ρ 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).The result of the learning over different values of

the learning rate and the number of patches is shown in fig. S7. Note that PGD is here implemented in continuous time, while in this setting we have discretized time for the action domain of the value function, selected using the greedy policy, $t^* = \text{argmax}_t \hat{Q}^\tau(r,t)$. 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 τ of the estimate $\hat{\rho}_k^\tau$ (c.f. eq. (8)) of ρ. We note that PGD learns faster in all cases. To demonstrate the insensitivity of PGD to the state space representation, at $t = 5 \times 10^5$, 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 is

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

with $\beta = 1/(1+\tau)$ called the learning rate, and τ the characteristic time of the exponential window of the corresponding continuous time filter over which the history is averaged. Exceptionally, here t indexes absolute time rather than trial time. To leading order in β, $\hat{\rho}_t \approx \beta \sum_i^t R_i$, 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 $\hat{\rho}_t \approx \beta \sum_i^t R_i \to \rho$ when t is large.

The rewards in this task are sparse: $R_t = 0$ except when a trial ends and the 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 efficient. Resolving a geometric series leads to the cumulative update to [7, 27]

$$\hat{\rho}_k = (1-\beta)^{T_k}\hat{\rho}_{k-1} + (1-(1-\beta)^{T_k})\rho_{\text{trial},k} \ , \tag{8}$$

where the smoothed reward, $\rho_{\text{trial},k} = R_k/T_k$, can be interpreted as a trial-specific reward rate. The initial estimate, $\hat{\rho}_0$, is set to 0. Exceptionally, $\hat{\rho}_1 = R_1/T_1$, 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 τ and denote the filtered estimate from its application with a τ-superscript, $\hat{\rho}_k^\tau$. For example, the precision of $\hat{\rho}_k^{\tau_{\text{long}}}$ as an estimate of a stationary reward rate ρ is set by how many samples it averages over, which is determined by the effective length of its memory given by τ_{long}.

Tokens task: a random walk formulation

The tokens task is a continuing task of episodes (here trials), each presenting 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 $N_0 = 0$. A fixed $t_{\max}$ number of jumps are realized with a duration between jumps of 200ms. 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 (Θ 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 $r = \Theta(N^2_{t_{\max}}) = 1$. The task was designed to study gain-optimal, ie. reward rate maximizing policies, rather than those that maximize accuracy. In particular, the task has additional structure that allows for controlling what the gain policy is through the incentive to decide early, α, 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_{ITI}, to make suboptimal the strategy of predicting randomly at the trial's beginning. We emphasize that it is through the trial duration that α serves as a task parameter controlling the strength of the incentive to decide early. When α is fixed, the corresponding reward rate maximizing policy that gives the optimal stationary reward rate, ρ_α. This policy shifts from deciding late to deciding early as α is varied from 0 to 1 (c.f. fig. S8j,k).

We consider a version of the task where α 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 α dynamics are responsive (maybe even adversarial), whereas the within-trial random walk dynamics (controlled by the rightward 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, $\bar{r}(\boldsymbol{S}_t)$.

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 and a proposal for how to learn the expected reward for the general task class we consider.

A $t_{\max}$-length sequence of random binary variables form a realization of a finite spin chain, $\boldsymbol{S}_{t_{\max}} = (S_1, \ldots, S_{t_{\max}})$, $S_t = \pm 1$, $i = 1, 2, \ldots, t_{\max}$. Consider a simple case in which each is an independent and identically distributed Bernoulli sample, $P(s) = p^{\frac{1+s}{2}}(1-p)^{\frac{1-s}{2}}$. 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, $\xi := \text{sgn}(N_{t_{\max}}) \in \{+, -\}$ 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 $P(\boldsymbol{s})$. Marginalizing (here just integrating out) the additional degrees of freedom leads to a binomial distribution in the number of + symbols, $N_t^+ = \sum_{i=1}^{t} \Theta(s_i) = (t + N_t)/2$,

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

with $n \in \{0, \ldots, t\}$ 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 ξ, 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 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 [15], 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,

$$\begin{aligned} \langle r | N_t = n \rangle &= \mathbb{E}\left[\Theta(N_{t_{\max}} N_t) | N_t = n\right] &(16)\\ &= \max\{p_{n,t}^+, 1 - p_{n,t}^+\} \,. &(17) \end{aligned}$$

The regret $r_{\max} - \langle r | N_t = n \rangle$, then also evolves on a lattice (see fig. 3(b)).

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

$$p_{n,t}^+ \approx \frac{1}{1 + \exp\left[-(at + b)n\right]} \tag{18}$$

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 accuracy is less than 0.05. A similar approximation without time dependence was presented in [15]. We nevertheless used the exact expression eq. (15) in all calculations.

PGD implementation and fitting to relaxation after context switches

In order to highlight its specific features, we sought to compare the PGD algorithm to data without additional model components. While behavioural noise certainly exists and would require adding stochastic sources to the model, we believed PGD was sufficient to account for the trajectory of ensemble average behaviour. To test this, we implemented PGD in its minimal form of only two parameters: τ_{long} and τ_{context}. These are the time constants of reward filters estimating the stationary reward rate, ρ, and the context-specific performance, ρ_{α_k}, for the trial context sequence, $\boldsymbol{\alpha}_k$, respectively. To fit these parameters to data in a way that still allowed us to validate the model on other, more informative features of the data, we used as a target for the fit the relaxation dynamics of trial-averaged decision times around switches in context.

Specifically, we identified the times of these 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, the policy is defined as 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 $t = t_{\text{dec}}$. Conditioning this ensemble on the state gives $P(t = t_{\text{dec}}|N_t, t) = p(N_t, t = t_{\text{dec}}|t)/p(N_t|t)$. To reduce estimator variance, we focus on the corresponding survival function, $P(t < t_{\text{dec}}|N_t, t)$. So, $P(t < t_{\text{dec}}|N_t, t) = 1$ 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 modulus the time step shown in fig. S10.

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 [6], 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 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 opportunity cost of time. 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 [10], 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 π thereafter. The action set for these binary decision tasks consist 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,

$$Q(s, a|t) = \mathbb{E}^{\pi}\left[\sum_{t'=t}^{T} R_t - \rho \,\middle|\, S_t = s, A_t = a\right] + \mathbb{E}^{\pi}\left[V(s|T+1)|S_t = s, A_t = a\right] \tag{19}$$

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. The expectation is over all randomness conditioned on following the policy, π, which itself could be stochastic. 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. (19) can be fixed so that $V_0 = 0$. Also, note that when the trial sequence is correlated, e.g. with context, $V(s|T + 1) \neq V(s|t = 0)$.

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 integrate over the remainder of the trial

since no further decision is made and so depend on the value at the start of the following trial. The resulting optimality equation for the value function $V(s|t)$ is then

$$\begin{aligned} V(s|t) &= \max_a Q(s,a|t) \ , \\ Q(s,\pm|t) &= r(s,\pm|t) - C(t) + V(s|t=T+1) \ , \\ Q(s,wait|t) &= -c(t) + \mathbb{E}_{s_{t+1}|s}\left[V(s_{t+1}|t+1)\right] \ , \\ V(s|t=0) &= V(s|t=T+1) \ . \end{aligned} \tag{20}$$

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_{ITI}, T satisfies $T_{ITI} \leq T \leq t_{\max} + T_{ITI}$. $C(t)$ is the portion of trial cost incurred after the decision, and $c(t)$ is the cost rate at time t. In general then, $C(t) = \sum_{t'=t+1}^{T} c(t')$. 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. (20) 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. (20) [6] 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 $t_{\max}$ 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 treated in [6]. Of course, ρ is unknown *a priori*. Within the dynamic programming approach, its value can be found in practise by exploiting the self-consistency constraint that the final value obtained by the recursion in the method, $V(s|t=0) = V(s|t=T+1)$. This dependence can be seen formally by taking the state-action value eq. (19), choosing a according to π to obtain the state value, $V(s|t)$, and evaluating it for $t = 0$,

$$V(s|t=0) = \mathbb{E}_{t_{\rm dec}}\left[\sum_{t=0}^{T} R_t - \rho\right] + V(s|t=T+1) \tag{21}$$

$$= \mathbb{E}_{t_{\rm dec}}\left[r(t_{\rm dec}) - \rho T(t_{\rm dec})\right] + V(s|t=T+1) \tag{22}$$

$$= \bar{R} - \rho\bar{T} + V(s|t=T+1) \ . \tag{23}$$

Here, $\bar{R} = \mathbb{E}_{t_{\rm dec}}[R]$ and $\bar{T} = \mathbb{E}_{t_{\rm dec}}[T]$ denotes the expectations over the trial ensemble that, when given the state sequence, transforms to an average over $t_{\rm dec}$, 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 self-consistency on eq. (23) gives $\rho = \bar{R}/\bar{T}$.

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_{track} and T_{sys} be the timescale of tracking and of the 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_{track}: the slower the timescale of tracking, the lower the cognitive cost. For simplicity, we set $q = 1/T_{\text{track}}$. 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 $\hat{\rho}_k^{\tau_{\text{context}}}$, τ_{context}. We redefine τ_{context} as

$$\tau_{\text{context}} \leftarrow \frac{\tau_{\text{context}}}{1 + Q^{\nu}} , \tag{24}$$

where ν 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. A natural choice for T_{sys} is T_k, the trial duration. For T_{track}, we introduce the filtered estimate of the trial duration, $\hat{T}_k^{\tau_{\text{context}}}$ (c.f. eq. (8)). Thus, the tracking timescale adapts to the system timescale. As a result of how τ_{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 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

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

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 in either the actions or the states.

Using an additional parameter γ, 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. S9. For $\gamma > 0$ the upper component shifts up proportional to γ. For $\gamma < 0$ the lower component 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.

ACKNOWLEDGMENTS

M.P.T. would like to acknowledge helpful discussions with Jan Drugowitsch, Zach Kilpatrick, Becket Ebitz, Paul Masset, and Anne Churchland.

---

[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," (2020), 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] Nils Kolling and Thomas Akam, "(Reinforcement?) Learning to forage optimally," Current Opinion in Neurobiology **46**, 162–169 (2017).

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

[6] 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).

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

[8] 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).

[9] 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).

[10] 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.

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

[12] 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).

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

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

[15] 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).

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

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

[18] 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).

[19] 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 , 2020.10.22.350280 (2020).

[20] 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).

[21] 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).

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

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

[24] 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).

[25] 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).

[26] Benjamin Y Hayden and Yael Niv, "The case against economic values in the orbitofrontal cortex (or anywhere else in the brain)," PsyArXiv , 1–26 (2020).

[27] 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.

[28] D. Thura. Personal communication.

[29] 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).

[30] 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).

[31] 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 LP – 956 (2016).

[32] 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.

[33] Iva K Brunec and Ida Momennejad, "Predictive Representations in Hippocampal and Prefrontal Hierarchies," bioRxiv , 786434 (2020).

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

[35] 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).

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

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

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

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

[40] 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).

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

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

[43] 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).

[44] Payam Piray and Nathaniel D Daw, "A simple model for learning in volatile environments," PLOS Computational Biology **16**, e1007963 (2020).

[45] 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.

[46] William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle, "Hyperbolic Discounting and Learning over Multiple Horizons," (2019), arXiv:1902.06865 [stat.ML].

[47] 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).

[48] Personal communication, Thomas Thierry.

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

[50] 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).

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

Supplemental Materials

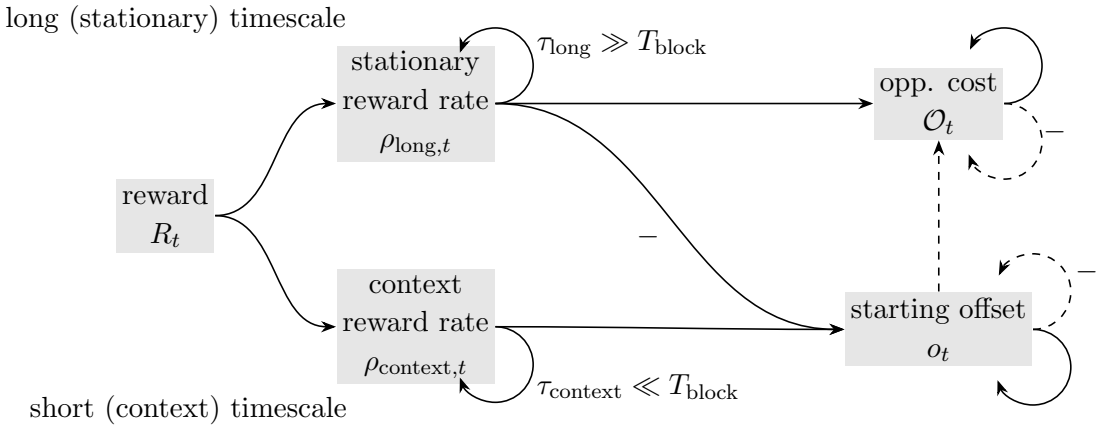

Figure S1. *Reward filtering scheme for online computation of within-trial opportunity cost.* With t denoting absolute time, the reward sequence, R_t, is integrated on both a stationary (τ_{long}) and context (τ_{context}) filtering timescale to produce estimates of the stationary and context-specific reward rates, respectively. These are large and small, respectively, relative to the average context switching timescale, T_{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 $\mathcal{O}_t$ is set to o_t, and the latter is zeroed. Thus, the opportunity cost starts at this offset and then integrates ρ_{long}, $\mathcal{O}_{t,k} = o_{T_{k-1},k-1} + \rho_{\text{long},k-1}t$, 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.

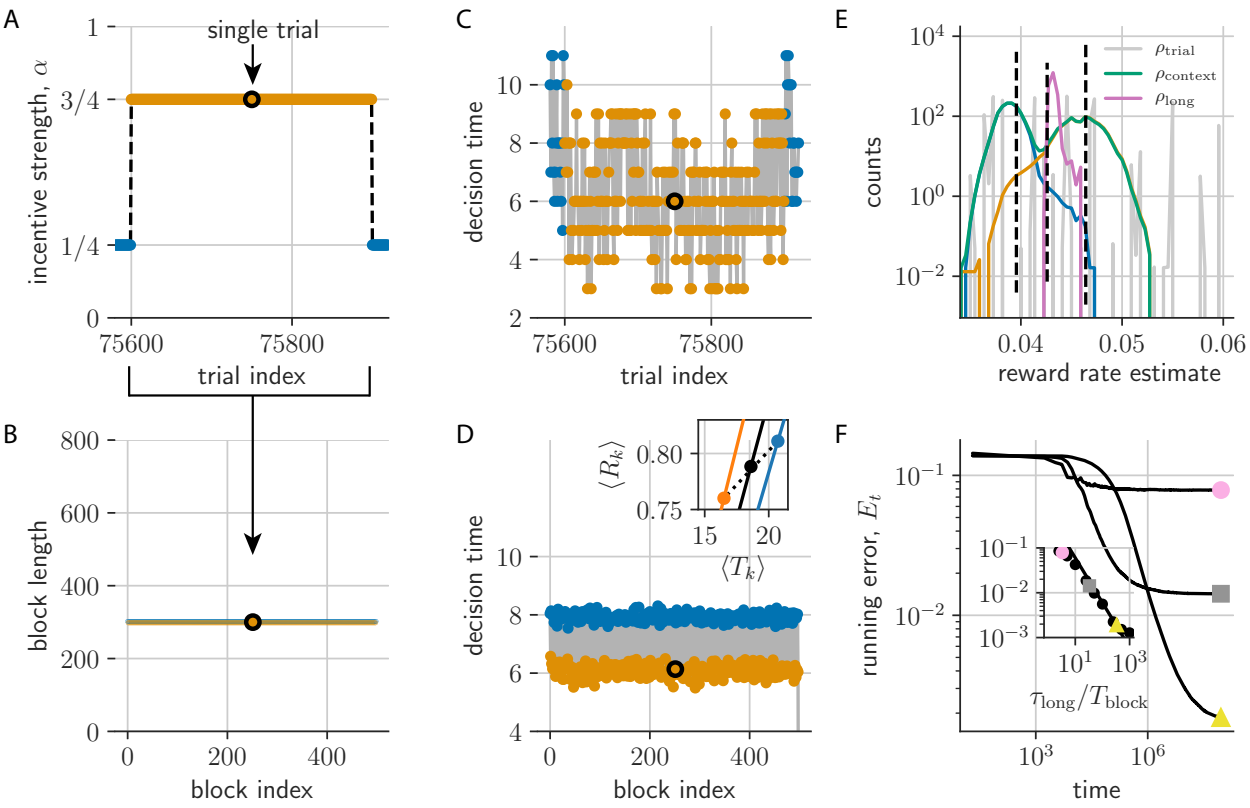

Figure S2. *PGD agent plays the tokens task with periodic α-dynamics.* (a) Trials are grouped into alternating trial blocks of constant α (fast (orange) and slow (blue) conditions). (b) Here, trial block durations are constant over the experiment. (c) Decision times over the trials from (a) distribute 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, ρ_{trial} (gray). The conditioned averages of $\hat{\rho}_k^{\tau_{\text{context}}}$ shown as blue and orange. (f) The relative error in estimating ρ, $E_t = \frac{1}{t}\sum_k^t |\hat{\rho}_k^{\tau_{\text{long}}} - \rho|/\rho$, 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 τ_{long} and T_{block} as expected (black line).

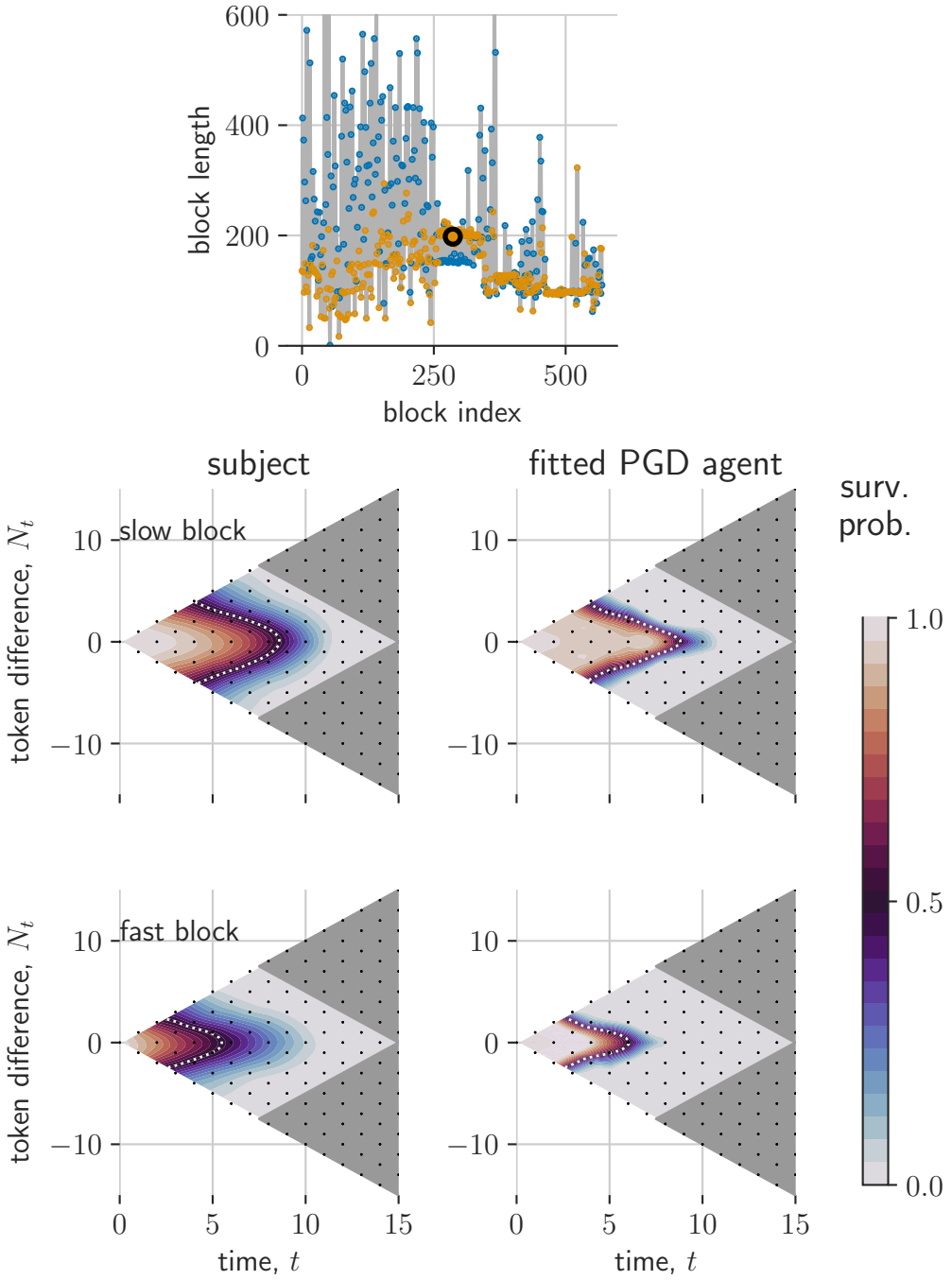

Figure S3. *Comparison of PGD and NHP in non-stationary α dynamics from [S18]: Subject 2.* Same as fig. 5.

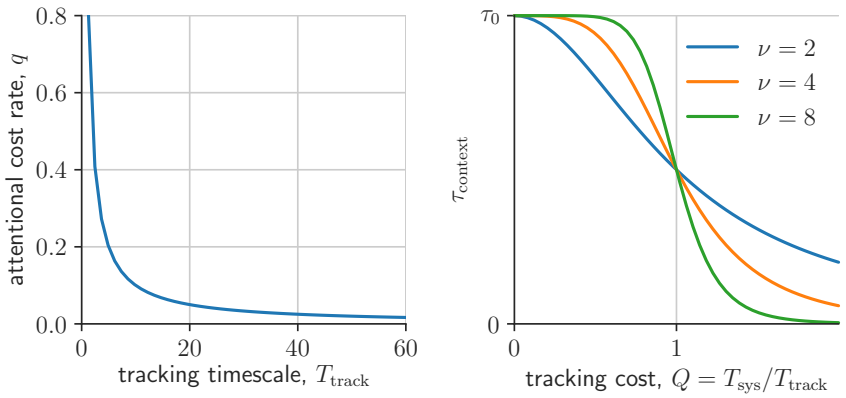

Figure S4. *Asymmetric switching cost model.* (a) Attentional cost rate, q, is set to be inversely proportional to tracking timescale, T_{track}. (b) Filtering timescale τ_{context} is scaled down with tracking cost, $Q = T_{\text{sys}}/T_{\text{track}}$ from a base timescale, here denoted τ_0 (shown for three values of sensitivity $\nu = 2, 4, 8$).

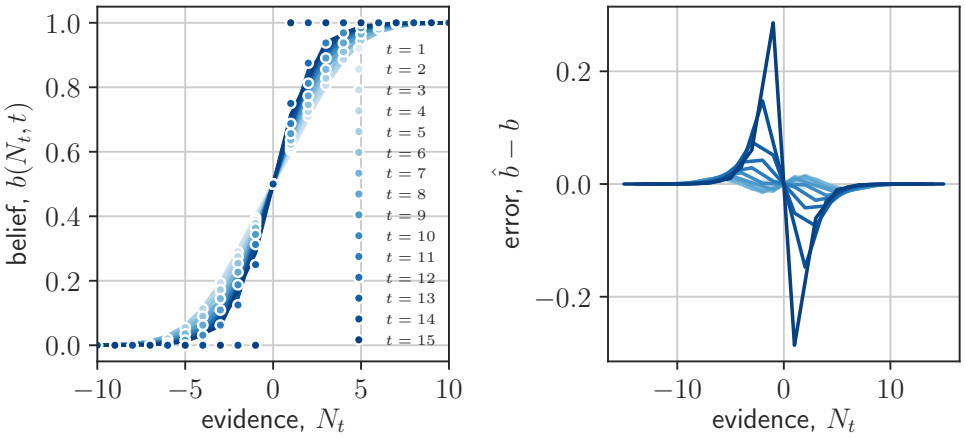

Figure S5. *sigmoidal approximation to expected reward.* (a) the approximation explained in Methods: State-conditioned expected trial reward, for different decision times. (b) The error in the approximation for different decision times.

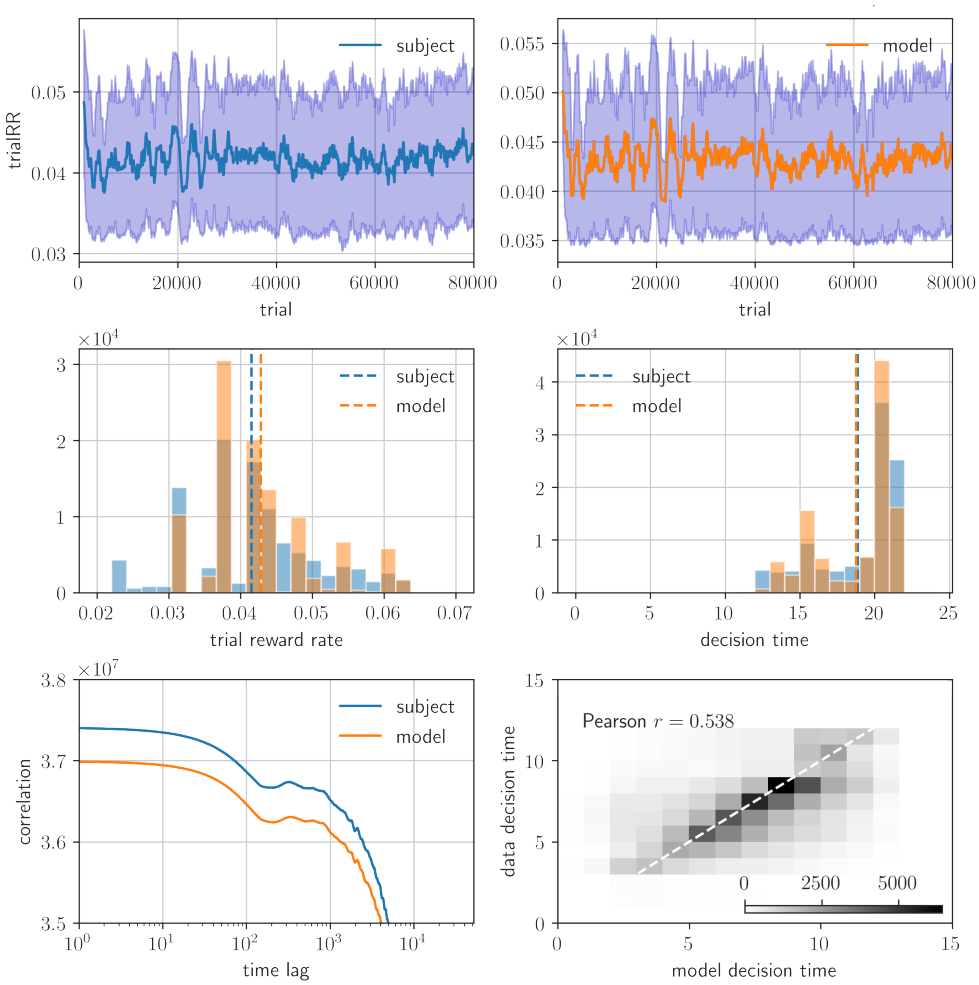

Figure S6. *Model validation on behavioural statistics from [S18].* Top: Running average trial reward rate $\rho_{\text{trial},k}$ over 1000 last trials. Middle: distributions of trial reward rate (left) and decision time (right). Bottom: Auto-correlation functions (left) and cross-correlation (right: gray-scale is trial count; white dashed line is perfect correlation)

.

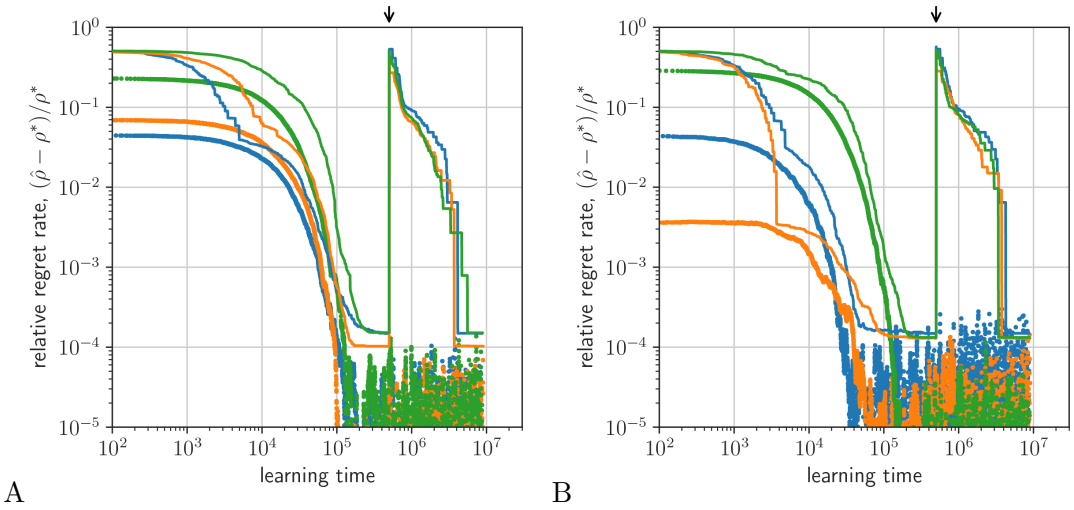

Figure S7. *Comparison of PGD and AR-RL learning on a patch leaving task.* Performance is defined as relative regret rate, $(\hat{\rho}-\rho^*)/\rho^*$ (PGD (dots); RL (lines)). (a) Performance over different sizes of the state vector ($d = 100$ (blue), 200 (orange), 300 (green)). (b) Performance over different learning rates (parametrized by integration time constant, $\tau = 1 \times 10^4$ (blue), 2×10^4 (orange), 3×10^4 (green)). (parameters: $\lambda = 1/5$; $r_{\max}$ sampled uniformily on $[0, 1]$). A random state label permutation is made at the time indicated by the black arrow.

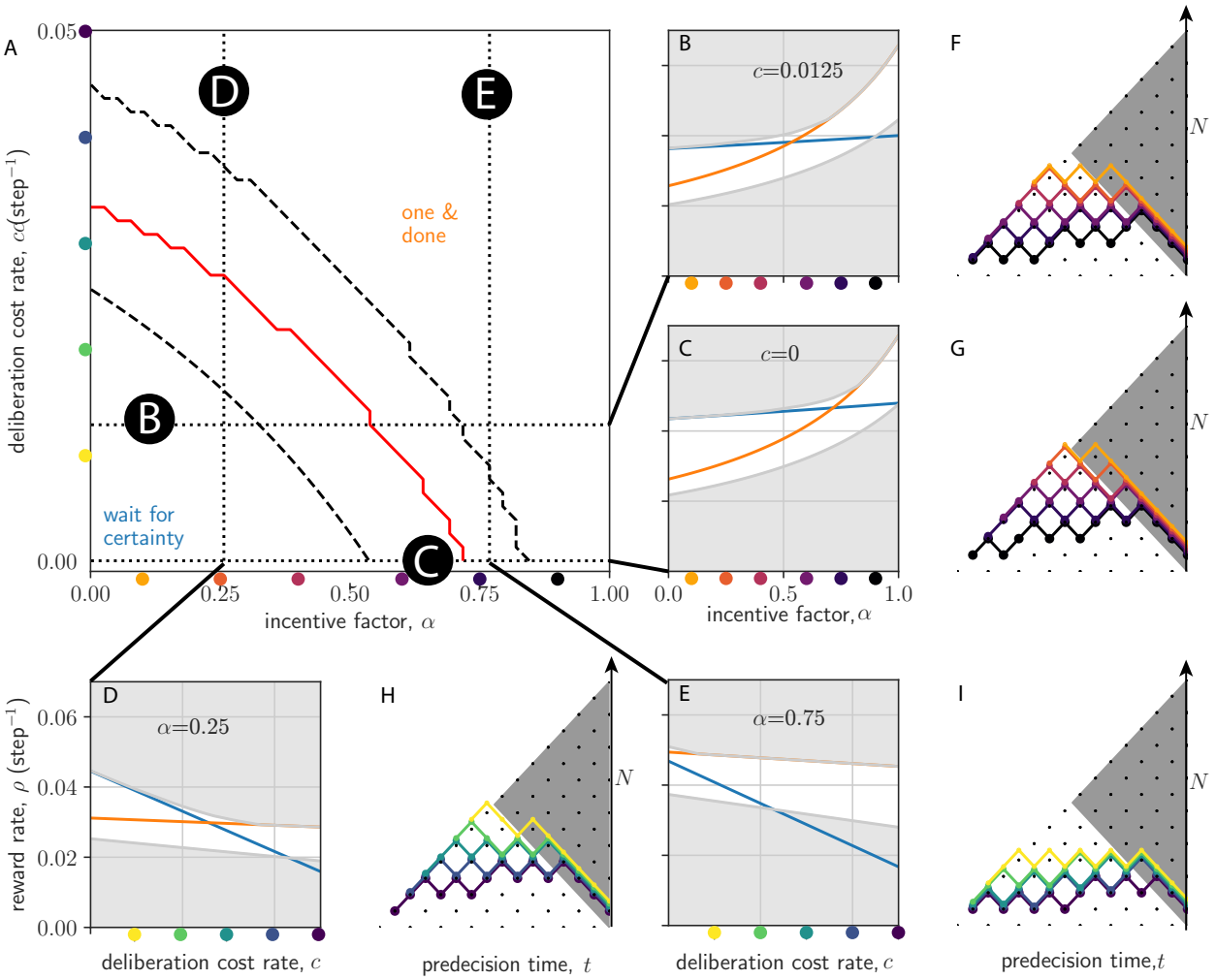

Figure S8. *Reward rate optimal strategies in* (α, c) *plane.* (a) The reward-rate maximizing policy interpolates from the wait-for-certainty strategy at weak incentive (low α) and low deliberation cost (low c), to the one-and-done strategy at strong incentive (high α) and high deliberation cost (high 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 (α, c)-plane. Shown are the reward rate as a function of α (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

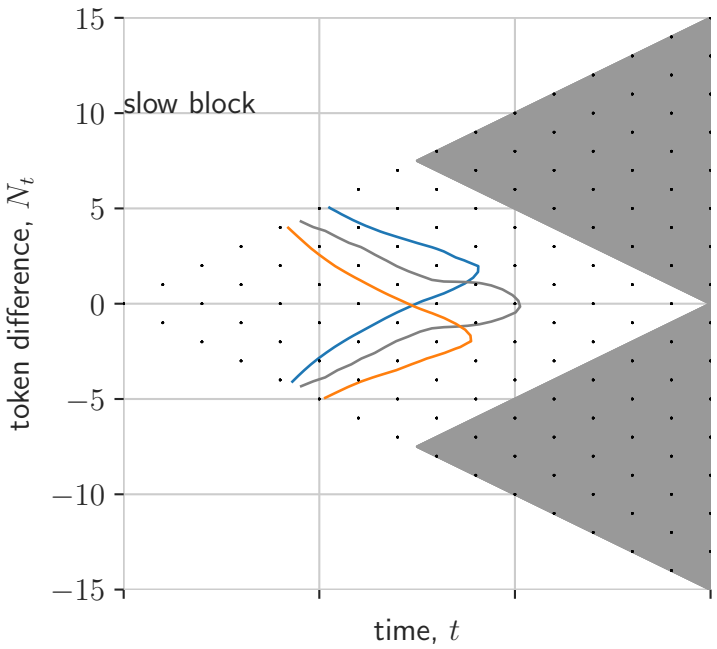

Figure S9. *Asymmetric action rewards skew survival probability.* Here, we plot the half-maximum of the PGD survival probability for three values of the action reward bias, $\gamma = -0.6, 0, 0.6$ (blue, black and orange, respectively). Other model parameters same as in fitted model.

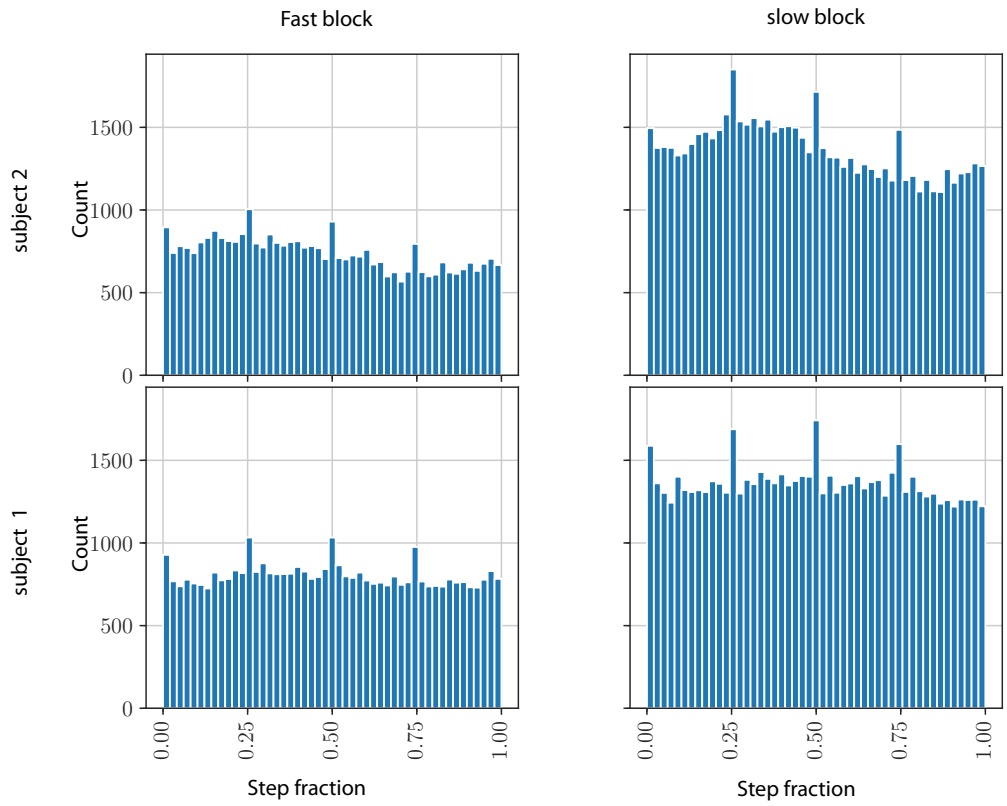

Figure S10. *Decision times relative to token jumps.* Here, we plot the histograms of decision times using their position between token jumps, the step fraction. The data is separated by α and monkey.