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

Maximilian Puelma Touzel, Paul Cisek, and Guillaume Lajoie

Abstract

The value we place on our time impacts what we decide to do with it. Value it too little, and we can obsessively attend to all details. Value it too much, and we carelessly rush to move on. How to strike this often context-specific balance when quantifying opportunity cost at each instant 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 generalize average-reward reinforcement learning to handle non-stationary contextual factors using a structured opportunity cost inferred over multiple time-scales. We then show that a heuristic solution that directly uses this context-aware cost readily adapts to changes in context. We apply these results in a hypothesis—well-grounded in both cognitive and systems decision-making neuroscience—for how this heuristic is implemented in primate nervous systems. Our proposal formally identifies neural urgency as the direct neural correlate of this non-stationary opportunity cost, prescribing its features like the baseline and initial slope based on performance history. We use behaviour and neural recordings from non-human primates in a non-stationary random walk prediction task to verify our results and make readily testable predictions for both neural activity and behaviour.

INTRODUCTION

Humans and other animals make a wide range of decisions throughout their daily lives. Any given act usually involves different decisions made at multiple levels and 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 other things an agent could potentially do with it. Estimating time's value is not straightforward for a number of reasons. It requires 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 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. Finally, an agent may not have full knowledge of these contextual factors and their stability or volatility. The agent's sensitivity to the risk associated with this uncertainty will then skew the value it assigns to possible alternatives.

These are significant, practical complications of making decisions contingent on *opportunity costs* [1], the formal economic concept capturing the value of the alternative activities lost by committing a scarce resource to a given use. Nevertheless, the opportunity cost of time was identified with the reference used in relative definitions of value, most notably the average reward used in average-reward reinforcement learning (AR-RL) [2, 3]. AR-RL has since been used to explain human and animal behaviour in foraging [4], free-operant conditioning [3], perceptual decision-making [5, 6], cognitive effort/control [6, 7], and even economic exchange [8]. Up to now, however, this theory and most of its applications have been for stationary average reward, ie. for fixed context, 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 de-

sign. We present subjects with specific choices in separate trials but without the option to just leave the lab and do something else. 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. Consequently, what subjects do within a given trial 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 relative value decision-making under uncertainty in a setting relevant to decision-making neuroscience. Using the theory, we develop a simple heuristic strategy called Performance-gated deliberation (PGD). 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. This heuristic effectively implements a collapsing decision boundary in probabilistic decision-making [5] and thus links to its putative neural correlate, "urgency" [5, 9–11]. These features arise in policies designed around improving reward rate rather than more classical concepts of fixed accuracy criteria [12]. However, PGD neither feeds the opportunity cost into an optimization over a model of the task statistics as in AR-RL [5], nor extracts urgency from data [13]. Instead, PGD uses the opportunity cost directly as urgency in a well-motivated approximation of AR-RL for adjusting the decision policy as a function of the agent's experience.

To illustrate how PGD applies in a specific decision-making scenario, and to make explicit links to neural mechanisms, we analyzed behavioral data collected over eight years from two non-human primates performing the "tokens task", a probabilistic guessing task in which sensory information about the correct choice is continuously changing within each trial, and the incentive to decide early (the context) is varied over longer timescales. Behavior in the task, in both humans [10] and monkeys [13], is consistent with a consensus about how neural dynamics underlying time-sensitive decision-making is implemented in different tasks and decision-making brains areas. In particular, neural recordings in monkeys suggest that evidence is estimated in dorsolateral prefrontal cortex [14], a growing context-dependent urgency signal is provided by the basal ganglia [15], and the two are combined to bias [16] and time [17], respectively, a competition between potential actions that unfolds in dorsal premotor and primary motor cortex [13, 18]. We provide a theoretical explanation for why decision-making mechanisms should be organized in this way and for how the brain can independently learn its evidence and urgency signals to achieve a good return on time investment by balancing between immediate rewards and the cost of time across multiple timescales.

RESULTS

A. Theory of performance-gated deliberation

1. Opportunity cost, regret, and a drawback of average-reward reinforcement learning

We consider tasks consisting of a long sequence of trials indexed by $k = 1, 2, \ldots$ (see fig. 1a). In each trial, a trial time t-indexed sequence, $\boldsymbol{s}_t = (s_0, \ldots, s_t)$, of states, s_t, is observed that provide evidence for an evolving belief about the correct choice. The chosen option and time, $t_{\mathrm{dec},k}$, to report determine both the reward received, R_k, and the

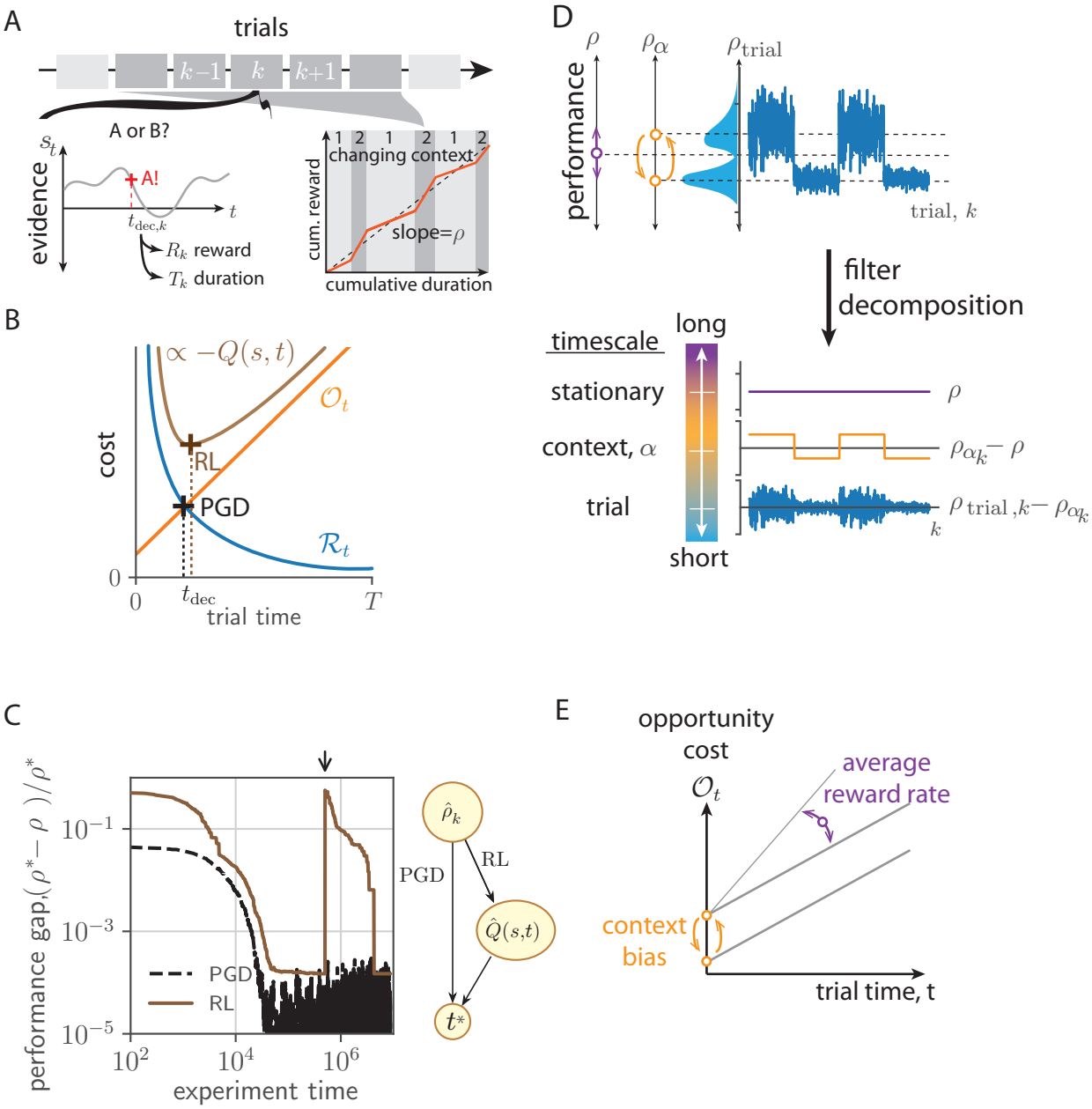

Figure 1. *Performance-gated deliberation.* (a) Task setting. Left: Within trial evidence, s_t evolves over trial time t in successive trials indexed by k. A decision (e.g. 'A') is reported at the decision time (red cross), determining reward and trial duration. Right: Reward rate (slope of red line), varies with context 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 $\mathcal{O}_t + \mathcal{R}_t \propto -Q(s,t)$, with $Q(s,t)$ the action value function. The PGD rule (black cross) finds t at which they intersect, $\mathcal{O}_t = \mathcal{R}_t$. (c) Convergence of performance towards optimum over learning in a patch-leaving task. Left: Shown is the performance gap: how much less is the current ensemble averaged reward rate, ρ, compared with the optimal reward rate ρ^* (RL: brown, PGD: black). The arrow indicates when the state labels were randomly permuted. Right: Schematic diagram of each algorithm's dependency. The AR-RL algorithm in addition estimates a value function. (d) Trial, context, and effectively stationary timescales in an experiment form a hierarchy of components to reward history, decomposed by filtering. (e) The corresponding trial opportunity cost grows with slope ρ, and is offset by the context deviation, $(\rho_\alpha - \rho)T_\alpha$.

trial duration, $T_k \geq t_{\text{dec},k}$. For a fixed strategy, the *stationary reward rate* is

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

Free-operant conditioning, patch leaving, and perceptual decision-making tasks often fall into this class. Previous work [5, 19] has studied the choice probability $p(R = r|\boldsymbol{s}_t, t_{\text{dec}} = t)$, for which the belief of correct report $p(R = 1|\boldsymbol{s}_t, t_{\text{dec}} = t)$ equals the expected trial reward $\bar{r}(\boldsymbol{s}_t, t)$ for binary rewards [5] (see [20] for further relation between value-based and perceptual decisions). We have suppressed conditioning these quantities on the choice reported since we will not explicitly address the problem of exploration and so assume the choice reported is always the one with the largest belief at decision time.

The *decision regret* at time t within a trial is given by the difference,

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

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. Average-reward reinforcement learning (AR-RL), first developed in artificial intelligence [21], was identified as accounting for this cost [3] and subsequently used to understand reward-based decision-making in neuroscience and psychology [5, 6, 22, 23]. Through use of the average-adjusted return $\sum_{t'>t}(R_{t'} - \rho)$, AR-RL aims for the highest ρ possible. The average-adjusted trial return can be expressed $r_{\max} - (\mathcal{R}_t + \mathcal{O}_t)$, where $\mathcal{O}_t$ is the opportunity cost, such that it is maximized by jointly minimizing $\mathcal{O}_t$ and $\mathcal{R}_t$ (fig. 1b). From this perspective, an agent's solution to the speed-accuracy trade-off is about how it balances decaying regret and growing opportunity cost. For $T_k = t_{\text{dec},k}$, AR-RL uses $\mathcal{O}_t = \rho t$.

We argue that value and return representations are a liability in real world tasks where task parameters can vary because they have as many entries as the number of states multiplied by the number of actions, which can be large, limiting the speed at which these representations adapt. Consider the following patch foraging task. An animal feeds among a fixed set of food (e.g. berry) patches. Total berries consumed in a time t saturates in time according to some given profile, shared across patches, as the fewer berries left are harder to find. Patches differ in their richness (e.g. berry density), randomly sampled and fixed over the task. Indexing patches with s, the food return is analogous to $\bar{r}(s, t)$, 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 more details). During its experience, the animal can estimate the average-adjusted trial return and leave at its maximum, the AR-RL solution. In fig. 1c, we show how performance improves towards the optimum ρ^* as the estimation of the AR-RL trial return improves with experience. However, if the environment undergoes a significant perturbation, e.g. a plant disease randomly lowers yields (applied as a random state label permutation at the indicated time), the performance of this AR-RL algorithm is essentially back to where it started. This weakness also afflicts approaches that directly learn policies instead [24]. Could high-value decision times be obtained without having to store value over context?

2. Performance-Gated Deliberation

We propose that instead of the maximum operation at the center of the AR-RL optimal solution (equivalent to the minimim of $\mathcal{O}_t + \mathcal{R}_t$), the agent simply take the intersection of $\mathcal{O}_t$ and $\mathcal{R}_t$ (shown as the black cross in fig. 1b).

$$t_{\text{dec}} := \min_t \{t \mid \mathcal{O}_t \geq \mathcal{R}_t\} \quad \text{(PGD decision rule)} \tag{3}$$

We call this heuristic rule at the center of our results *performance-gated deliberation* (PGD). Plotted alongside the AR-RL performance in fig. 1c, PGD achieves better performance than AR-RL overall and is completely insensitive to the perturbation. It achieves this without a value function at all, by basing its decision of when to decide solely on $\mathcal{O}_t$ and $\mathcal{R}_t$. We note that we constructed the patch return profile in this toy example such that PGD is the AR-RL optimal solution. In general, however, PGD will be sub-optimal, which is why we call it a heuristic. We also used the fact that the animal could directly observe the reward. In the more general stochastic setting, the animal will have to learn the state associations in the expected reward, $\bar{r}(s,t)$, over the residual uncertainty in the trial. However, this lower-level learning is more likely to be stationary across trials, independent of context, and we study such a case here. Finally, the non-stationarity in this toy example was a perturbation at a single time point. In general, task non-stationarity will be extended in time and thus a broader notion of opportunity cost that accounts for multiple timescales is needed.

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

The stationary reward ρ is estimated to high precision as $\hat{\rho}_k^{\tau_{\text{long}}}$ from applying a low pass filter with a long integration time, τ_{long}, to the reward sequence R_k [6, 25]. To model changing context, we introduce a context parameter, α, to the task that varies across trials but has a stationary distribution so that ρ is still well-defined (e.g. seasonal context is negligible when averaging over many years). The stationary performance if α were fixed is denoted ρ_α. For a simple non-stationary setting, we consider a sequence, α_k, that varies on a single timescale, τ_{context}, set by, for example, a deterministic periodicity or the switching rates of a Markov process. We assume that τ_{context} has been learned and will infer it from data when we later analyze animal experiments. Estimating performance on this timescale leads to an estimate, $\hat{\rho}_k^{\tau_{\text{context}}}$, that is intended to vary with time, unlike $\hat{\rho}_k^{\tau_{\text{long}}}$, and tracks the effective instantaneous, context-specific performance ρ_{α_k}.

$\hat{\rho}_k^{\tau_{\text{context}}}$ at first appears like straightforward way to extend the AR-RL formulation of stationary opportunity costs to the non-stationary case, using $\mathcal{O}_t = \hat{\rho}_k^{\tau_{\text{context}}} t$. However, if τ_{context} can be learned, then the agent has the opportunity to plan on these timescales. Such context-aware plans are naturally structured as a so-called *decision hierarchy*: sequences of moment-to-moment actions are grouped into plans executed in particular context. The opportunity costs of these plans are distinct from those incurred by the moment-to-moment actions contained within them: they are incurred on distinct timescales. Using $\mathcal{O}_t = \hat{\rho}_k^{\tau_{\text{context}}} t$ incorrectly lumps them together in a time average incurred moment-by-moment.

We propose a multiple timescale decomposition of performance as an effective decision hierarchy from which opportunity costs at distinct levels are, by construction, incurred in a self-consistent manner. In this decomposition, ρ serves as the opportunity cost rate component associated with moment-by-moment decisions at the base of the hierarchy. Additional

components of the opportunity cost arise from conditioning on knowledge about, and thus the ability to plan on timescales beyond the moment-to-moment over which the task varies. Each component adds a zero time-average conditional variation at the respective timescale to the sum of components below it (see fig. 1d). In practise, the actual decomposition used will rely on the model that the agent deploys about its own performance, such that the averages are over beliefs. As with ρ, each modelled component must be be estimated by the agent. Only in the case where the agent incorporates the belief that the task lacks variation beyond moment-by-moment does the average opportunity cost reduce to ρ, the stationary average-reward, and conventional AR-RL is recovered. If instead knowledge of context variation in the form of τ_{context} is incorporated, the opportunity cost becomes

$$\mathcal{O}_t = \rho t + (\rho_\alpha - \rho)T_\alpha \text{ (context-aware opportunity cost) .} \tag{4}$$

Here, the 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. (4) 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. 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. using $\rho \equiv \sum_\alpha \rho_\alpha T_\alpha / \sum_\alpha T_\alpha$ when trials from different context are uniformily sampled). 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. (4) from reward history.

B. Neuroscience applications: PGD in the tokens task

We applied the PGD algorithm to the tokens task. Briefly, the subject must report its 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. Importantly, after the subject reports, the interval between jumps contracts to once every 50ms ("fast block") or once every 150ms ("slow block"), giving the subject the possibility to save time by taking an early guess. The contraction factor, $\alpha = 3/4$ and $\alpha = 1/4$ respectively, quantifies the incentive strength to decide early, which serves as the task context, α (see Methods for details; fig. 2a). There are thus many within trial states and the state dynamics is stochastic, in contrast to the patch leaving task example used above. The expected decision regret (computed in Methods) evolves on a lattice, always starting at 0.5 and ending at 0 (see fig. 2d). We assume the agent has learned to track this decision regret. A means to do so is outlined in the discussion.

1. PGD in a passive non-stationary tokens task

We first show the behaviour of the algorithm in the simple case where α switches back and forth every 300 trials (see fig. 2b,c). We set τ_{context} at a few tens of trials and τ_{long} so that it averages over many blocks. In this case, the PGD agent model, i.e. its decision boundary and the quantities that determine it, relaxes into a noisy periodic trajectory, over a fast and slow block (fig. 2g). The decision times relax after a context switch (fig. 2e) to their

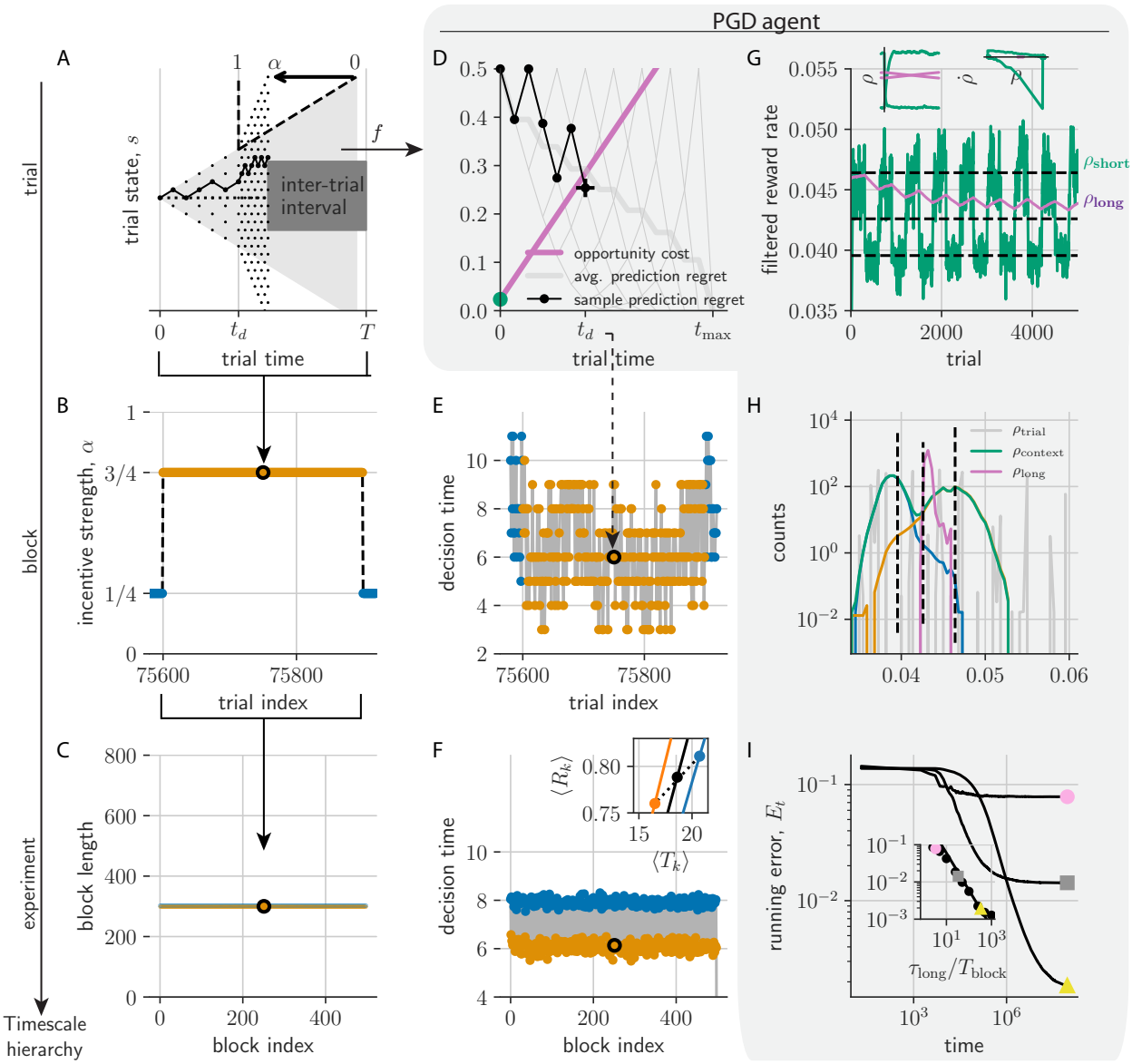

Figure 2. *PGD agent plays the tokens task with periodic α-dynamics.* (a) A tokens task trial for $\alpha = 3/4$ and decision time t_d. (b) Trials are grouped into alternating trial blocks of constant α (fast (orange) and slow (blue) conditions). (c) Here, trial block durations are constant over the experiment. (d) Decision space obtained from (a) through the transformation, f, from evidence to regret, $\mathcal{R}_t$. All expected decision regret trajectories (gray lattice; thick gray: trial-averaged) start at 0.5 and ends at 0. The one from (a) is shown in black. t_d is determined by the crossing of the regret and opportunity cost (purple). (e) Decision times over the trials from (b) distribute widely, but relax after context switches. (f) 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). (g) Expected rewards filtered on τ_{long} (ρ_{long}, purple) and τ_{context} (ρ_{context}, green). Insets show their dynamics (left: in time; right: in phase space) over each of the two blocks. Black dashed lines from bottom to top are $\rho_{\alpha=1/4}$, ρ, and$\rho_{\alpha=3/4}$. (h) Distribution of estimates have lower variance than the trial reward rates, ρ_{trial} (gray). The conditioned averages of ρ_{context} shown as blue and orange. (i) The relative error in estimating ρ, $E_t = \frac{1}{t}\sum_k^t |\rho_{\text{long}} - \rho|/\rho$, 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).

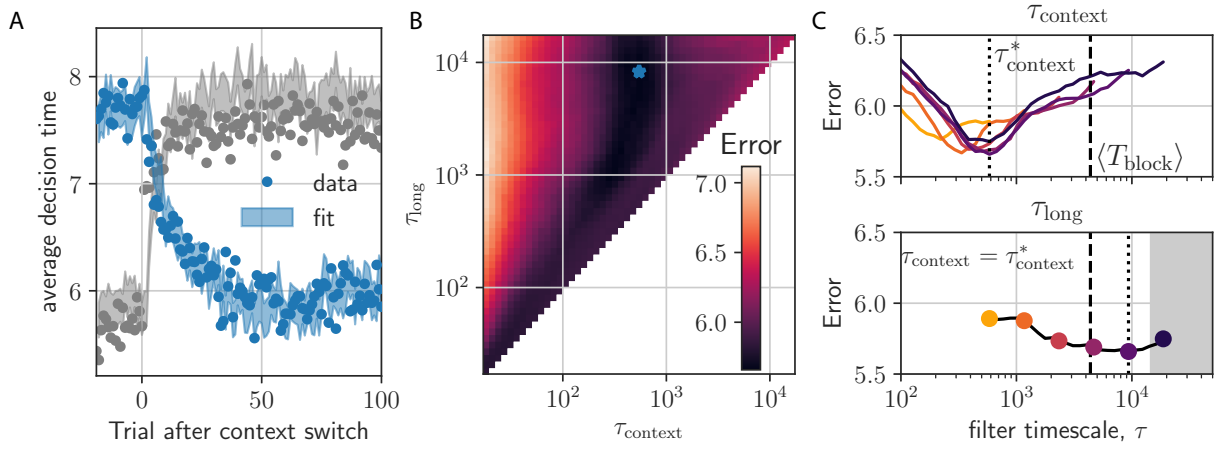

Figure 3. *Model fit.* (a) Average decision times (dots) aligned by context switch. Shaded region is standard error of model, which includes the asymmetric switching component described in the Methods. Model fitted only to slow-fast transition (blue), not fast-slow transition (gray). (b) Error of model fit over τ_{context} and τ_{long}. (c) Cross sections of (b) at fixed τ_{long} (top) and τ_{context} (bottom).

conditional average but exhibit strong fluctuations from the random sequence of random walk realizations. The block average decision times vary little over blocks of the same type (fig. 2f). The PGD algorithm sacrifices accuracy in the fast context to achieve shorter trial duration and achieves a higher context-conditioned reward rate compared to decisions in the slow block (the slopes shown in the inset of fig. 2f). The resulting estimates ρ_{long} and ρ_{context}, are near their stationary values (dashed lines in fig. 2g,h). While these estimates are more precise for larger integration windows (larger τ_{context} and τ_{long}, respectively), they nevertheless exhibit some bias (fig. 2h), as a result of the residual zigzag relaxation over the period of the limit cycle. When the block duration, T_{block}, is much less than τ_{long}, the within-block exponential relaxation is roughly linear and so the average unsigned deviation between ρ_{long} and the actual stationary reward, ρ, is $1 - \exp\left[-T_{\text{block}}/\tau_{\text{long}}\right] \approx T_{\text{block}}/\tau_{\text{long}}$. This scaling fits the data well (fig. 2i: inset). On the other hand, if $T_{\text{block}}/\tau_{\text{long}}$ is large, ρ_{long} approaches ρ_{short} and opportunity cost undergoes strong baseline shifts at context switches. These are transient, however, and most of the time the opportunity cost is given by the first component in eq. (4), with the context specific reward rate as the slope. We propose this limit as a test of the theory in the discussion.

2. PGD in an active non-stationary tokens task and its comparison to behaviour by non-human primates

Next, we applied the PGD algorithm to the actual α sequence used in the experiments reported in [18] (see fig. 4a). Trials were structured in blocks, but in contrast to the above example, with large, irregular variations in size. To fit the model parameters, τ_{context} and τ_{long}, we looked to the animal's decision time statistics at context switches (see fig. 3a). The slow-to-fast transitions happened smoothly on average, whereas the fast-to-slow switch happening almost instantaneously. We thus focussed on fitting only slow-to-fast transition.

While not necessary for this fit, we developed a simple tracking-cost model that accounts for the asymmetric relaxation times with addition of a single tracking-cost sensitivity parameter (see Methods; fig. 3a). Applying the fitting procedure precisely identified τ_{context}, but only set a lower bound on the value of τ_{long} (fig. 3b,c). The resulting behavioural statistics for these fitted parameters gave good correspondence with the data (see fig. S5).

The state and time conditioned survival probability (i.e. that a decision has not yet occurred) of the action policy associated with a stationary strategy makes for a robust and rich representation of the behavioural statistics (see Methods;fig. 4b-e). The absence of behavioural noise sources in the model means it underestimates the variability. Nevertheless, the remarkably smooth average strategy is well captured by the model. Fast block strategies are shifted earlier by similar amounts relative to slow block strategies in both model and data. Note that the model has essentially only a single degree of freedom (τ_{context}), which we have fit using a timescale extrinsic to the within trial decisions.

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 optimal solutions for stationary contexts using average-adjusted value functions optimized using dynamic programming (see Methods for a value function formulation for this class of tasks that builds off of previous work [5]). We included a constant deliberation cost rate, c [5], incurred during the deliberation period in each trial. The reward rate as a function of the incentive strength, α is shown in fig. 4f for $c = 0$. The optimal solution interpolates from the wait-for-certainty strategy at low α to the one-and-done strategy [26] at high α (this holds also with c and thus over the entire (α, c)-plane; see fig. S8). The performance of the α-conditioned reward rates achieved by the two primates and a reference human are also shown. Naturally, 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 model fit, this intermediate performance is shared by the clearly suboptimal PGD agent. Importantly, the family of optimal strategies across c for the slow ($\alpha = 1/4$) and fast ($\alpha = 3/4$) block conditions are qualitatively distinct from the data (fig. 4g,h). Optimal policies shift around the edges of the relevant decision space, while the policy extracted from the data lies squarely in the bulk. Optimal strategies thus invariably wait to resolve any remaining ambiguity, while the primate and PGD behaviour resolve residual ambiguity at intermediate trial times (fig. 4b-e).

3. Neural urgency and opportunity cost

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. In fig. 5a, we restate in a schematic diagram the consensus understanding of the neural dynamics of decision-making using both the regret and opportunity cost formulation developed here and the equivalent belief space dynamics [5]. In the latter, the belief about being correct $p(r = 1|\boldsymbol{s}_t, t_{\text{dec}} = t)$ (denoted b_t here for brevity) rises towards a collapsing decision boundary. For purposes of comparison, we express this boundary as $C - u_t$, where an C is initial level of desired confidence that is lowered by a growing function of trial time $u_t > 0$. The decision rule is then $b_t > C - u_t$. In AR-RL optimal policies, u_t has a complicated dependence on opportunity cost. In PGD, C is interpreted as the maximum reward $r_{\max}$ and u_t is identically the opportunity cost.

We test our prediction that neural urgency reflects our formulation of a dynamic opportunity cost. With linear encoding (c.f. [14]), we assume a 1D choice manifold on which the

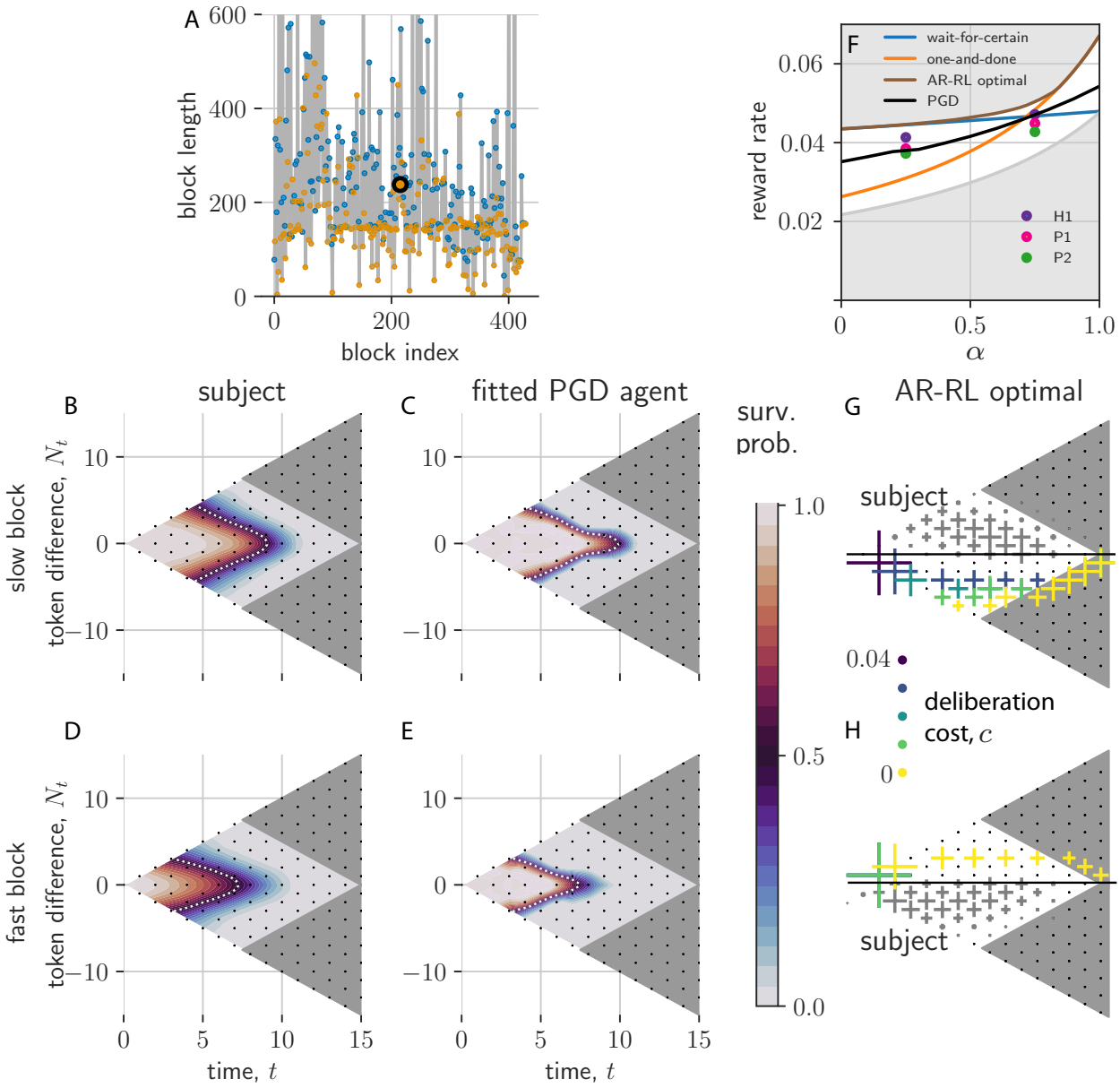

Figure 4. *Comparison of PGD and NHP in non-stationary α dynamics from [18].* (a) block length sequence used in experiment (c.f. fig. 2c). (b-e) Interpolated state-conditioned survival probabilities, $P(t_{\text{dec}} = t|N_t, t)$, over slow (b,c) and fast (d,e) blocks. White dashed lines show the $P(t_{\text{dec}} = t|N_t, t) = 0.5$ contour. (f) Shown are the reward rate as a function of incentive strength and no deliberation cost ($c = 0$) (wait-for-certainty shown in blue; one-and-done shown in orange). We additionally show the context-conditioned reward rates for the two primates (P1,P2) as well as a reference human (H1), and the PGD algorithm (black line). Reward rates for primates are squarely in between the best and uniformly random strategy (lines bounding the upper and lower gray regions, respectively). (g,h) Survival probability histograms from optimal decision boundaries across different values of the deliberation cost for slow (g) and fast (h) conditions. Only samples with $N_t < 0$ are shown to make room for the primate's histogram shown in gray. Cross size corresponds to histogram count. Note that, unlike primate data, all optimal strategies give no intermediate decision times at ambiguous ($N_t \approx 0$) states..

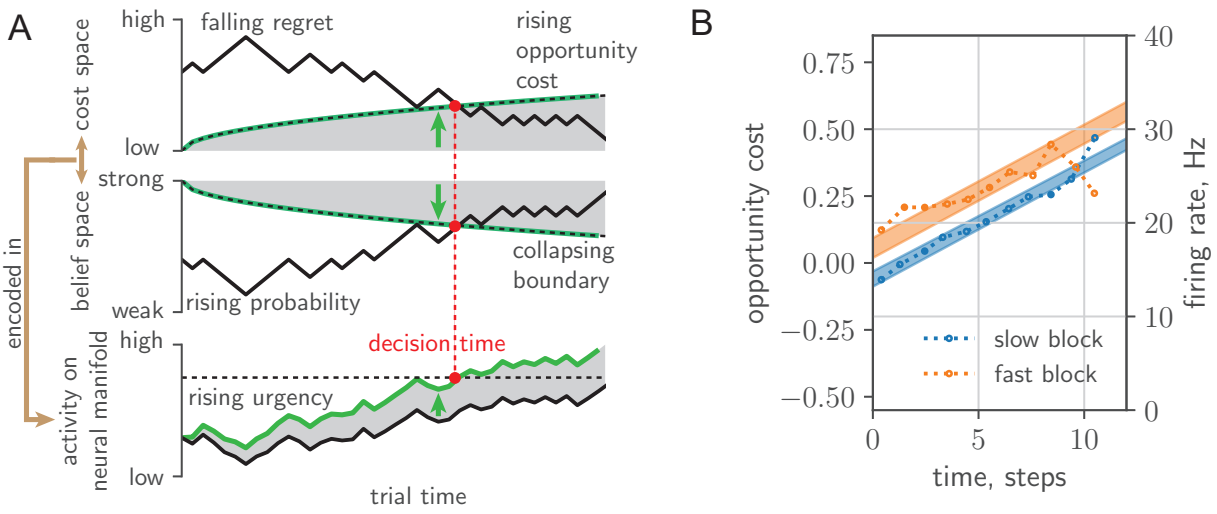

Figure 5. *Neural urgency and collapsing decision boundaries and comparison with data.* (a) Top: Falling regret (black) meets rising opportunity cost (green). Middle: Rising probability of success (black) meets collapsing decision boundary (green) in belief space. Bottom: Belief/regret is encoded (black) into a low-dimensional neural manifold, with the addition of an urgency signal (green). (c.f. fig.8 in [5].) (b) Opportunity cost linearly maps onto the urgency signal extracted from zero-evidence conditioned cell-averaged firing rate in PMd (c.f. fig.2f [18]).

sum of the encoded belief and urgency, $\tilde{b}_t + \tilde{u}_t$, drives the motor response once sufficiently large (e.g. $\tilde{b}_t + \tilde{u}_t > \tilde{C}$). A single neuron's contribution to $\tilde{b}_t$ will depend on its choice selectivity, while $\tilde{u}_t$ is a signal shared across all neurons. Thus, we can reveal $\tilde{u}_t$ by conditioning on zero-evidence ($\tilde{b}_t = 0$) and averaging over cells [18].

In fig. 5b, we show the population firing rate of the decision-making area (here PMd) conditioned on zero-evidence environment states over all trials from the data in [18]. With an unit conversion from reward to spikes/step (here simply distinct y-axes), the uncertainty in the estimated fast and slow context-conditioned opportunity cost signals tightly bracket the observed urgency signals. The unit conversion is roughly 0.03 units of cost per spike, where cost is in units of the reward delivered on correct trials. There are multiple features of the qualitative correspondence exhibited in fig. 5b: (1) the linear rise in time; (2) the same slope across the two conditions; (3) the offset between conditions and its order: the fast condition is offset to higher values than the slow condition. Each of these three features has a specific meaning now by interpreting urgency as opportunity cost: (1) the animal uses a constant step-wise opportunity cost rate, (2) this cost rate refers to moment-to-moment decisions that are agnostic of context, and (3) context-aware planning leads to a baseline opportunity cost with sign given by the deviation from the time-average of the context-conditioned reward rate. Thus, the neural activity driving context-conditioned behavioural responses is consistent with being gated by opportunity cost at the neural level, with earlier responses in high reward rate contexts because the opportunity cost of deliberation there is higher.

DISCUSSION

We have proposed a heuristic decision-making algorithm that gates deliberation based on performance. We gave a foraging example in which it is optimal with respect to the average-adjusted value function of average-reward reinforcement learning (AR-RL). In general, PGD will be suboptimal. It is however generally applicable and at once exploits the stationarity of the environment statistics while simultaneously hedging against longer term non-stationarity in reward context, striking a balance between strategy complexity and return. 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. Consequently, 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 representations [27].

There are both scientific and clinical implications of this proposal of a cortico-basal ganglia system in which performance on multiple, behaviourally-relevant timescales are broadcast to multiple decision-making areas to gate the speed of thier respective attractor-based decision-making dynamics [16, 28]. While the view that tonic dopamine encodes average reward is decades old, the multiple timescale representation has received increasing empirical support in recent years, from cognitive results [29–31] to a recent unified view of dopamine signalling of temporal difference errors using multiple discount factors [32] and of dopamine as encoding both value and uncertainty [33]. While the relationship of dopamine to time perception, and thus putatively to decision speed has been proposed in the reward learning literature [34], a proposal that includes how speed is implicated in the decision-making areas driving motor responses has been absent. Our theory goes a step further by considering dynamic evidence tasks and mechanistically identifying urgency as the means by which the reward code ultimately affects decisions. PGD is also an example of decision-making without an explicit value function, contributing to a contemporary debate about their necessity [24]. 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 divergent 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 [35–37]. We even provided a concrete proposal for a signal filtering system that extracts a context-sensitive opportunity cost from a reward sequence (equivalently a temporal difference error sequence putatively encoded by dopamine).

Beyond the estimation of opportunity cost, we assumed that the animal had the means to form a model of the expected reward from which it could compute the within-trial decision regret. For the tokens task, in particular, there is direct evidence of encoding of expected reward in dorsal lateral prefrontal cortex [14]. For the general class of tasks we consider, a generic, neurally plausible means to learn the expected reward is via distributional value codes [33]. 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[38]. The expected reward at a chosen future time can be easily computed using this representation.

Our theory is prescriptive. For behaviour, this is via the shape of the action policy (c.f. fig. 4). PGD varies markedly with reward structure, and thus provides a wealth of predictions for this policy. For example, a salient feature of the standard tokens task is its reflective symmetry in 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 (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 remain fixed across blocks. One simple prediction is that the slope should exhibit increasing variation across block type with longer blocks.

Our work impacts modern reinforcement learning by describing how to generalize AR-RL to dynamic reference values based on beliefs. This epistemic perspective parallels that used to interpret the discount-reward formulation as a result of assumptions on the volatility of the environment [39]. This suggests a new class of algorithm between model-based and model-free that, being based in the average-reward rather than the discount-reward formulation, is perhaps better suited to the continuing task setting in which the conventional successor representation was developed. We have left a detailed algorithmic analysis of this to future work, but expect improvements, as with successor representations, in settings where it is advantageous to decouple the learning of environment statistics and the learning of reward structure.

In the space of strategies, PGD lies in regime between exploiting detailed knowledge (average-case optimal) and using conservative heuristics (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 [40], 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 [41], using it to further understand the computational advantages of PGD is an interesting direction for future work.

More generally, humans, despite our apparent access to sophisticated computation, still exhibit measurable bias in how we incorporate past experience [42]. One simple example is the win-stay/lose-shift strategy, a more rudimentary kind of performance-gated decision-making than PGD that explains how humans approach the rock-paper-scissors game [43]. In this 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 supports our 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 an analytically tractable patch leaving task for which PGD learning is optimal with respect to the average-adjusted value function (related but not equivalent to the marginal value optimum of optimal foraging, for which the decision rule is $\mathcal{O}_t > r_{\max} - \mathcal{R}_t = \bar{r}(s,t)$ [4]). Here the value is simply the return from the patch. This allowed us to compare PGD's convergence properties relative to conventional reinforcement learning algorithms that use value functions. In contrast to PGD, the latter in general require exploration. For

a setting generous to the RL algorithms, we allowed them 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 to substantially slower convergence.

In this task, the subject is randomly switched between 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, $p(r_{\max})$ over a range of positive values. The trials of the task are temporally extended periods during which the subject consumes the 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 property of the environment. The return diverges negatively for vanishing patch leaving times for mathematical convenience, but one could imagine situations where leaving a patch soon after arriving is prohibitively costly. A stationary policy is then a leaving time, t_s, for each of d patches. Given any policy, the stationary reward rate for uniformly random sampling of patches is then defined as

$$\rho = \sum_s^d r_s(t_s) \Big/ \sum_s^d t_s \text{ (patch average)}. \tag{5}$$

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 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 average-reward reinforcement learning. Given the return profile we have chosen, the corresponding optimal decision time in the sth patch, $t^* = \sqrt{r_{\max}/(\lambda\rho)}$, 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 $\mathcal{O}_t = \mathcal{R}_t$ here takes the form $\rho t = r_{\max}/(\lambda t)$). Thus, they share the same optimal reward rate, ρ^*. Using t^* for each patch in eq. (5) gives a self-consistency equation for ρ with solution $\rho^* = \mu_1^2/4\mu_{1/2}^2\tau$, where $\mu_n = \langle r_{\max}^n \rangle$ (we have assumed d is large here to remove dependence on the realization of the set of $r_{\max}$).

The result of the learning over different values of the learning timescale and the number of patches is shown in fig. S7. PGD is 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 a finite lower bound on the performance gap, i.e. the relative precision $\epsilon = (\rho^*-\rho)/\rho^* > 0$ for the AR-RL algorithm. Approaching this bound, both PGD and AR-RL learning convergence time is limited the integration time τ of the estimate $\hat{\rho}_k^\tau$ (c.f. eq. (7)) of ρ. We note that PGD learns faster in all cases. To demonstrate the insensitivity of PGD to the state space representation, at $t = 5\times10^5$, 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.

Filter methods

For unit steps of discrete time, the step-wise update is

$$\hat{\rho}_t = (1-\beta)\hat{\rho}_{t-1} + \beta R_t \ , \tag{6}$$

with $\beta = 1/(1+\tau)$ an effective parameter called the learning rate, and τ the characteristic time of the exponential window. Exceptionally, here t indexes absolute time rather than trial time. Note that for $\tau \sim \mathcal{O}(t) \gg 1$, $\beta \sim \mathcal{O}(1/t) \ll 1$ in which case $\hat{\rho}_t \approx \beta \sum_i^t R_i \to \rho$ when t is large. Also note that the rewards are sparse: $R_t = 0$ except when a trial ends and the trial reward R_k (1 or 0) is received. More efficient then is a cumulative update of eq. (6) that smooths the reward uniformly over the trial duration and is applied once at the end of each trial. This reduces to [6, 25]

$$\hat{\rho}_k = (1-\beta)^{T_k}\hat{\rho}_{k-1} + (1-(1-\beta)^{T_k})\rho_{\text{trial},k} \text{ (online estimate)}, \tag{7}$$

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. (7) 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 hereon denote the filtered estimate from its application with a τ-superscript. The precision of $\hat{\rho}_k^\tau$ as an estimate of a stationary reward rate ρ is set by the length of its memory given by τ. This precision becomes high for timescales, τ_{long}, set much longer than all other timescales in the problem (e.g. trial duration). We will hereon use τ_{long} to denote the timescale over which the agent chooses to estimate the stationary reward rate, ρ. This estimate is then denoted $\hat{\rho}_k^{\tau_{\text{long}}}$.

Tokens task: a random walk formulation

The tokens task is a continuing task of episodes. In each episode, an unbiased random walk, $\boldsymbol{N} = (N_0, \ldots, N_{t_{\max}})$ with $N_t = \{-t, \ldots, t\}$ and $N_0 = 0$ and of a fixed $t_{\max}$ number of jumps plays out (the duration between jumps, typically 200ms, is used as a natural unit of time). 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, $t_{\text{dec}} \in \{0, 1, \ldots, t_{\max}\}$. The decision-making task then only involves choosing when to decide. The subject then 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).

A greedy policy for this symmetric (unbiased) random walk can use the sign of the state at the decision time, $\text{sign}(N_{t_{\text{dec}}})$ (and randomly if $N_{t_{\text{dec}}} = 0$) as its prediction. An explicit action space beyond decision time is thus not necessary but 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, $a_{t_{\text{dec}}} = \pm 1$, after which actions are disabled and the prediction is stored in an augmented state 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: Episodic decision-making and dynamic programming solutions of value iteration)

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 the incentive to decide early. Namely, the remaining $t_{\max} - t_{\rm dec}$ jumps after $t_{\rm dec}$ occur faster with parameter, α: $\alpha = 0$ no speed up, $\alpha = 1$ infinite speed up (thus the α used in a given trial is only observed by the agent after its decision). In particular, the trial duration for deciding at time t in the trial is

$$T_\alpha(t) = t + (1 - \alpha)(t_{\max} - t) + T_{\rm ITI}, \tag{8}$$

where a dead time between episodes, $T_{\rm ITI}$, is added to make suboptimal the strategy of predicting randomly at the episode's beginning. We have added the subscript α to T_α in order to 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, π_α, gives optimal stationary reward rate, ρ_α. π_α 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, $\bar{r}(\boldsymbol{s}_t, t)$.

Expected trial reward for the tokens task

For the tokens task, we derived and used an exact expression for the expected reward. We derive that expression here as well as a simple approximation and a proposal for how to learn the expected reward for arbitrary tasks in the general task class we consider.

A $t_{\max}$-length sequence of random binary variables form a realization of a finite spin chain, $\vec{\sigma} = (\sigma_1, \ldots, \sigma_{t_{\max}})$, $\sigma_t = \pm 1$, $i = 1, 2, \ldots, t_{\max}$. Consider a simple case in which each is an independent and identically distributed Bernoulli sample, $P(\sigma) = p^{\frac{1+\sigma}{2}}(1-p)^{\frac{1-\sigma}{2}}$. We are interested in functions of this trajectory, namely the sign of $N_t = \sum_{i=1}^t \sigma_i$, for some $0 \leq t \leq t_{\max}$ and in particular the probability of $\mathrm{sgn}(N_{t_{\max}}) \in \{+, -\}$ given N_t (note that N_t is even if t is even and same with odd values). We will remove the case of no sign in $N_{t_{\max}}$ by choosing $t_{\max}$ to be odd, for simplicity. The distribution of $\vec{\sigma}$ is

$$P(\vec{\sigma}) = \prod_{i=1}^{t_{\max}} P(\sigma_i) \ . \tag{9}$$

First, consider predicting $\mathrm{sgn}(N_t)$ with no prior information. $-t \leq N_t \leq t$ appears directly in $P(\vec{\sigma})$. Integrating out the additional degrees of freedom leads to a binomial distribution in the number of + symbols, $N_t^+ = \sum_{i=1}^t \Theta(\sigma_i) = (t + N_t)/2$, with $N_t^+ = 0, \ldots, t$,

$$P(N_t^+) = \binom{t}{N_t^+} p^{N_t^+}(1-p)^{t-N_t^+} \ , \tag{10}$$

with $N_t = 2N_t^+ - t$. Thus, the probability that $N_t > 0$, i.e. $N_t^+ > t/2$, is

$$p_t^+ := \sum_{N_t^+=0}^{t} \binom{t}{N_t^+} p^{N_t^+}(1-p)^{t-N_t^+}\Theta(N_t) \ . \tag{11}$$

Now consider predicting $\mathrm{sgn}(N_{t_{\max}})$, given N_t. Define $t' = t_{\max} - t$ as the remaining time steps to the predicted time and $N_{t'} = \sum_{k=t+1}^{t_{\max}} \sigma_k$, i.e. the total count in the remaining part of the realization, and $N^+_{t'}$ similarly, then the probability of $N_{t_{\max}} = N_t + N_{t'} > 0$ is defined in the same way as p^+_t

$$p^+_{t_{\max}|t} := \sum_{N^+_{t'}=0}^{t'} \binom{t'}{N^+_{t'}} p^{N^+_{t'}} (1-p)^{t'-N^+_{t'}} \Theta(N_t + N_{t'}) \,. \tag{12}$$

We incorporate the $\Theta(N_t + N_{t'}) = \Theta(N^+_{t'} - N^+_t - t_{\max}/2)$ factor by changing the upper bound of the sum to $\min\{t', N^+_t + (t_{\max}-1)/2\}$. If the upper bound is t' then $p^+_{t_{\max}|t} = (1 - \mathrm{sgn}(N_t))/2 \in \{0, 1\}$, and also for larger times, since the sum over its domain is normalized. Otherwise, the upper bound is $N^+_t + (t_{\max}-1)/2$, and the distribution is

$$p^+_{t_{\max}|t} = \sum_{N^+_{t'}=0}^{N^+_t + (t_{\max}-1)/2} \binom{t'}{N^+_{t'}} p^{N^+_{t'}} (1-p)^{t'-N^+_{t'}} \,. \tag{13}$$

For odd $t_{\max}$, $p^-_{t_{\max}|t} = 1 - p^+_{t_{\max}|t}$. For the symmetric case, $p = 1/2$, we can without loss of generality focus on the subset of trajectories for which $\mathrm{sgn}(N_{t_{\max}}) = +$, and obtain

$$p^+_{t_{\max}|t} = \frac{1}{2^{t_{\max}-t-N_t}} \sum_{N^+_{t'}=0}^{N^+_t + (t_{\max}-1)/2} \binom{t_{\max}-t}{N^+_{t'}} \,, \tag{14}$$

when $N_t < \frac{t_{\max}+1}{2} - t$ and 1 otherwise.

For deciding at time t when the random walk state $N_t = n$, and where the expectation is over the remaining jumps in the trial, we reparametrize the above expression using n,

$$\langle r | N_t = n, t\rangle = \mathbb{E}\left[\Theta(N_{t_{\max}} N_t) | N_t = n, t\right] \tag{15}$$

$$= \max\{p^+_{n,t}, 1 - p^+_{n,t}\} \,, \tag{16}$$

where we apply $p^+_{t_{\max}|t}$ with n substituted for N_t,

$$p^+_{n,t} = \frac{1}{2^{t_{\max}-t-n}} \sum_{n^+_{t'}=0}^{n^+_t + (t_{\max}-1)/2} \binom{t_{\max}-t}{n^+_{t'}} \,, \tag{17}$$

is the conditional probability that $N_{t_{\max}} > 0$ conditioned on the current state $N_t = n$ and time t, and where $n^+_t = (n+t)/2$ is the observed number of positive jumps up to time t, and $n^+_{t'}$ is the unobserved number of positive jumps in the remaining $t_{\max} - t$ steps. The space of trajectories, i.e. of $\vec{\sigma}$, maps to a space of trajectories of $p^+_{t_{\max}|t}$ defined on an evolving lattice in belief space (see fig. 2(b)).

This function has a simple sigmoid approximation,

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

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. S4. Approximation error is worse at t near $t_{\max}$. More than 95% of decisions times across the policies occur before 12 time steps, where the approximation error in accuracy is less than 0.05.

Survival probabilities over the action policy

Behavioural analysis 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. 4). 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, $P(t < t_{\text{dec}}|N_t, t)$. $P(t < t_{\text{dec}}|N_t, t) = 1$ 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. Note that to simplify the analysis, we have binned decision times by step. We ensured that there was no statistical information in the response times between token steps **??**.

Episodic decision-making and dynamic programming solutions of value iteration

As a starting point to apply our theory to episodic tasks, here we generalize the mathematical notation and description of an existing AR-RL formulation and dynamic programming solution of the random dots task [5], a binary perceptual evidence accumulation task extensively studied in neuroscience. We connect this extended formulation to the concept of decision urgency. 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 state value function $V(s|t)$ in which the highest of the state-action values, $Q(s, a|t)$, is selected. These functions are conditioned on a given trial time, t, where $t = 0$ is the trial start time. $Q(s, a|t)$ is the same function described in detail in the previous section, with the addition that the trial structure requires that the decision time relative to the trial be made explicit. So, $Q(s, a|t)$ is the value function of selecting action a when in state s, at possible decision time t within a trial, and then following action policy π after t. 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, $R(s, a|t) = \mathbb{E}^{\pi}\left[\sum_{t'=0}^{T} R_{t'}\right]$ 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 $R(s, a|t)$ at the end of the trial.

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

contribution of the current trial,

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

where T is the (possibly stochastic) trial end time and $V(s|t=T+1)$ is the state value at the start of the following trial. 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 $V(s|t=0) = V(s|t=T+1) = V_0$ 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, $V(s|t=T+1) \neq V(s|t=0)$. We treat this case in the following section.

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) [5] 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 (the case of $\delta\rho_t = 0$ considered in the previous section), $c(t) = \rho$. This is the AR-RL case treated in [5]. 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, $V(s|t=0)$, is equal to $V(s|t=T+1)$. 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 $t = 0$,

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

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

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

Here, $\bar{x} = \mathbb{E}_{t_d}[x]$ denotes the expectation over the trial ensemble that, when given the state sequence, transforms to an average over t_d, the trial decision time, defined as when $V(s|t)$ achieves its maximum on the state sequence, S_t. $R(t) := \max_{a\in\{-,+\}} R(s_t, a|t)$ is the expected trial reward for deciding at t, with trial-averaged reward, $\bar{R}$. $\bar{T}$ is the trial-averaged duration of a trial. Imposing self-consistency on eq. (23) gives $\rho = \bar{R}/\bar{T}$.

The expected trial return at decision time is the argument of the trial-average in eq. (22), $R(t_d) - \rho T(t_d)$, where $-\rho T(t_d)$ is the corresponding opportunity cost incurred in the trial. This trial-level formulation of opportunity cost is consistent with the following time step-level formulation. The effective opportunity cost of committing time to some (possibly temporally-extended) action is the cost rate integrated over the time it takes to execute the action, which is taken to be the time until the next possible action. For the class of tasks considered here, deciding to delay reporting by one additional time step in a trial in order to accumulate another sample of evidence costs the decision-maker, ρ in reward. Delaying for t time steps then incurs a cost ρt. Deciding instead to report in a trial incurs a cost given by the cost rate integrated until the next possible decision time, which is at the start of the next trial. The cost thus integrates over the remaining time in the trial, $\rho(T(t_d) - t)$. This is precisely the AR-RL formulation where the value incorporates a cost of ρ incurred at each time step for a total cost over the trial of $\rho T(t_d)$. The above formulation, including the solution method, allows for context-specific opportunity cost rates by replacing ρ by its time-varying value, $\rho + \delta\rho_\alpha$ (c.f. **??**). It the trial sequence is independent and identically distributed, then the self-consistency criterion above is satisfied. However, if they follow some correlated dynamics, then the self-consistency constraint must be adapted to account for the residual value incurred after the current trial from conditioning on the state and action in it.

Asymmetric switching cost model

Here, we present a small extension to the performance tracking component of the PGD agent aimed at capturing the asymmetric relaxation timescales after context switches observed in the primate behaviour of [18]. The basic notion is that tracking a signal at a finer timescale should be more cognitively costly, so that adapting from faster to slower environments should happen quickly so as to not pay this cost unnecessarily, compared to slow to fast, where the increasing cost paid is always commensurate with precision earned. We now develop this formally (see fig. S3).

Let T_{track} and T_{sys} be the timescale of tracking and of the system, respectively. One way to view the mismatch ratio, $T_{\text{sys}}/T_{\text{track}}$, is via an attentional cost rate, c. c should decay with T_{track} and for simplicity we consider $c \propto 1/T_{\text{track}}$. The mismatch cost over a characteristic time of the system is then $C = cT_{\text{sys}} = T_{\text{sys}}/T_{\text{track}}$, the mismatch cost. We propose that the

mismatch enters the algorithm via a scale factor on the integration time of the reward filter for ρ_{context}, τ_{context}. Thus for a reference time constant τ_{ref}, we define

$$\tau_{\text{context}} = \frac{\tau_{\text{ref}}}{1 + C^{\nu}} \ , \tag{24}$$

where ν is a sensitivity parameter. $\nu > 1$ captures the nonlinear sensitivity to the mismatch cost. that the timescale used to integrate the reward rate deviation in the bias term of the opportunity cost, T_{context} tracks the trial duration T_k using timescale τ_{context}. Thus, we set $T_{\text{sys}} = T_k$ the trial duration and $T_{\text{track}} = T_{\text{context}}$. ν is then the single free parameter added to the model in this extension.

Prediction for asymmetric rewards

Given a payoff matrix, A, 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

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

where R_{sa} is the reward for reporting $a \in \{-,+\}$ in the trial realization leading to s, the sign of $N_T$an. 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, A was the identity matrix. In this case, and for all cases where A is a symmetric matrix, $A = A^\top$, an equivalent decision rule is to decide based on the sign of N_t. When A is not symmetric, however, this is no longer a valid substitute. We propose to add an asymmetry in either the actions or the states.

Using an additional parameter γ, we can add asymmetry via a bias for + actions that leaves the total reward unchanged by replacing the payoff matrix with

$$\begin{bmatrix}R_{++}(1+\gamma) & R_{+-}(1-\gamma)\\ R_{-+}(1+\gamma) & R_{--}(1-\gamma)\end{bmatrix} \ ,$$

The result for $\gamma = -0.6, 0, 0.6$ 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 $N_t \propto \pm\gamma$.

ACKNOWLEDGMENTS

M.P.T. would like to acknowledge helpful discussions with Jan Drugowitsch, Zach Kilpatrick, 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," , 1–41 (2020), arXiv:2010.08920 [cs.LG].

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

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

[5] Jan Drugowitsch, Anne K Churchland, Michael N Shadlen, and Alexandre Pouget, "The Cost of Accumulating Evidence in Perceptual Decision Making," **32**, 3612–3628 (2012).

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

[7] 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).

[8] 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).

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

[10] Paul Cisek, Aude Puskas, and Stephany El-murr, "Decisions in Changing Conditions : The Urgency-Gating Model," **29**, 11560–11571 (2009).

[11] Anne K Churchland, Roozbeh Kiani, and Michael N Shadlen, "Decision-making with multiple alternatives," **11**, 693–703 (2008).

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

[13] 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).

[14] 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).

[15] 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).

[16] 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).

[17] B A J Reddi and R H S Carpenter, "The influence of urgency on decision time," (2000).

[18] David Thura, Guido Guberman, and Paul Cisek, "Trial-to-trial adjustments of speed-accuracy trade-offs in premotor and primary motor cortex," Journal of Neurophysiology **117**, 665–683 (2016).

[19] 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).

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

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

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

[23] 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).

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

[25] Nathaniel D Daw, "Chapter 16 - Advanced Reinforcement Learning," (Academic Press, San Diego, 2014) pp. 299–320.

[26] 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).

[27] 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).

[28] Kong-fatt Wong and Xiao-jing Wang, "A Recurrent Network Mechanism of Time Integration in Perceptual Decisions," **26**, 1314–1328 (2006).

[29] 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.

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

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

[32] 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).

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

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

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

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

[37] 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).

[38] 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.

[39] William Fedus, Marc G Bellemare, and Hugo Larochelle, "HYPERBOLIC DISCOUNTING AND LEARNING OVER MULTIPLE HORIZONS," **0** (2009), arXiv:arXiv:1902.06865v3.

[40] Personal communication, Thomas Thierry.

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

[42] 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).

[43] 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

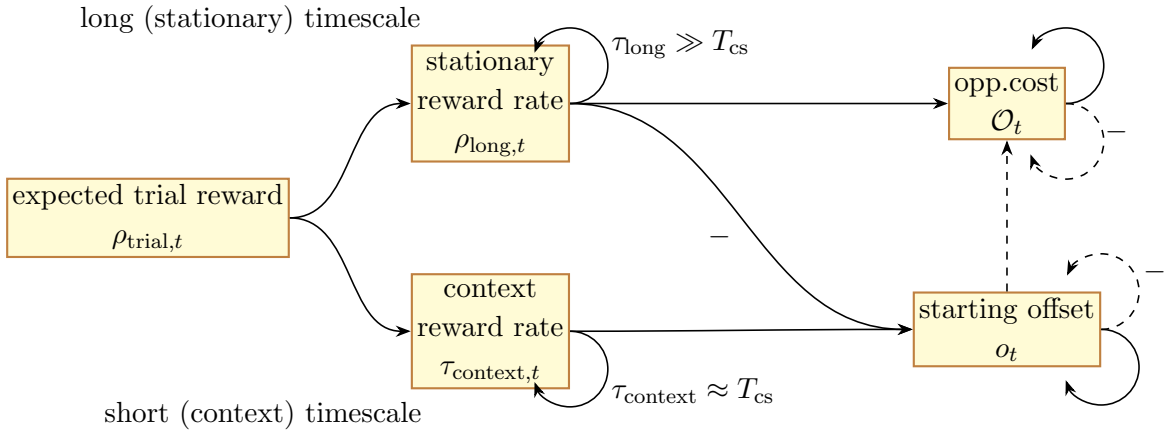

Figure S1. *Reward filtering scheme for online computation of within-trial opportunity cost.* The expected trial reward, ρ_{trial}, is integrated on both a stationary (τ_{long}) and context (τ_{context}) filtering timescale to produce estimated average and context-specific reward rate estimates, respectively. These are relative to the average context switching timescale, T_{cs}. The estimate of the context-specific offset, o_t is computed by integrating the difference of these two estimates. When a trial terminates, its value is added to the opportunity at the same time that $\mathcal{O}_t$ and o_t are 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.

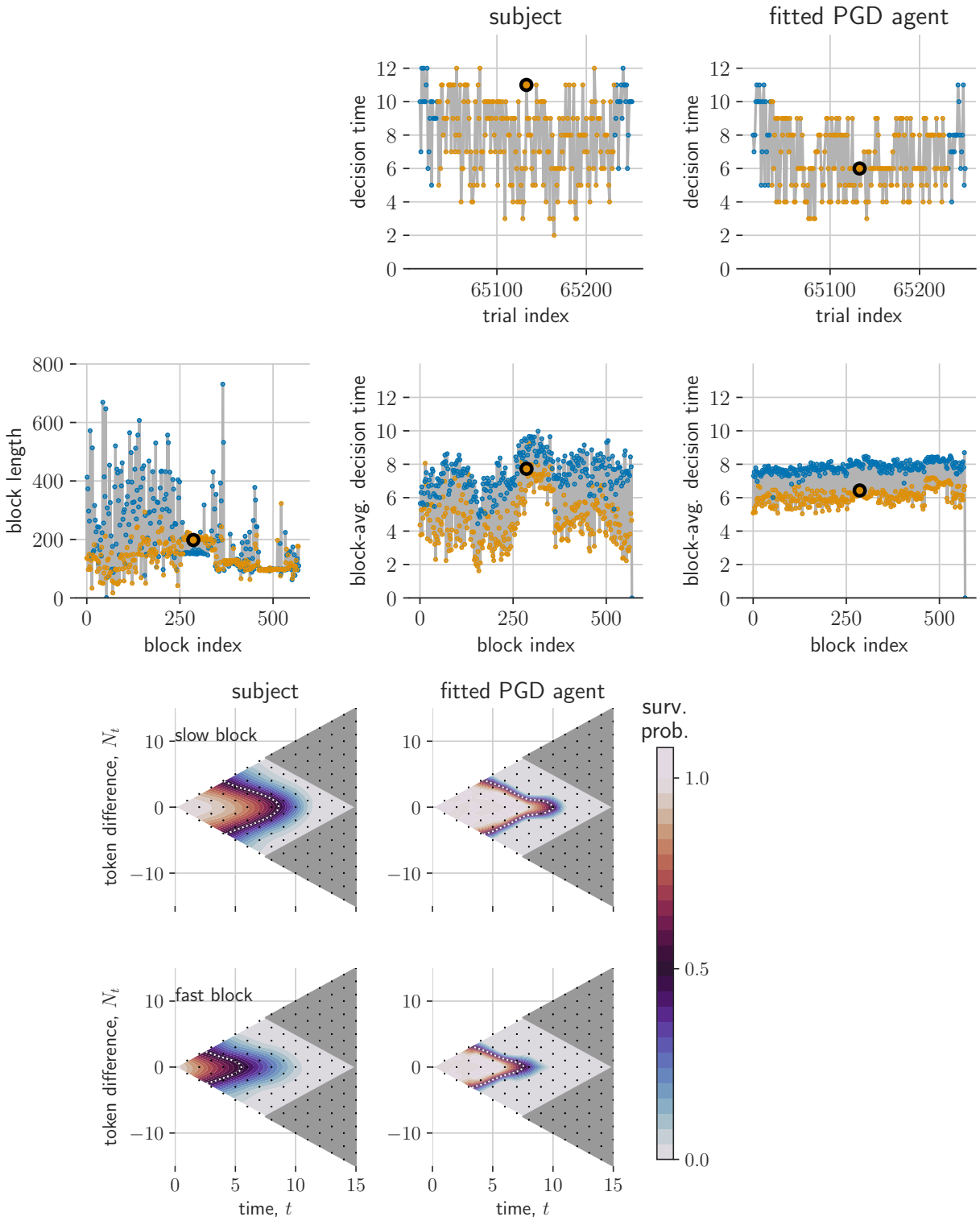

Figure S2. *Comparison of PGD and NHP in non-stationary α dynamics from [S18]: Subject 2.* Same as fig. 4.

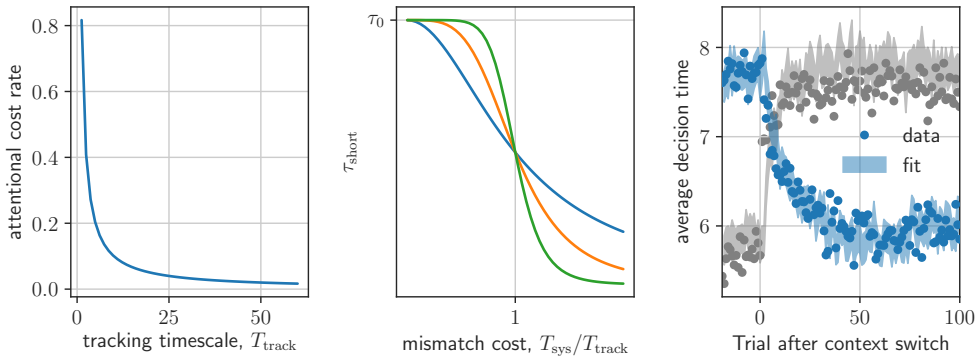

Figure S3. *Asymmetric switching cost model.* (a) Cost rate is inversely proportional to tracking timescale, T_{track}. (b) Filtering timescale τ_{context} scales down with mismatch cost $T_{\text{sys}}/T_{\text{track}}$ (sensitivity $\nu = 2, 4, 8$). (c) Adding this modified τ_{context} gives good fits to both types of context switches ($\nu = 9$).

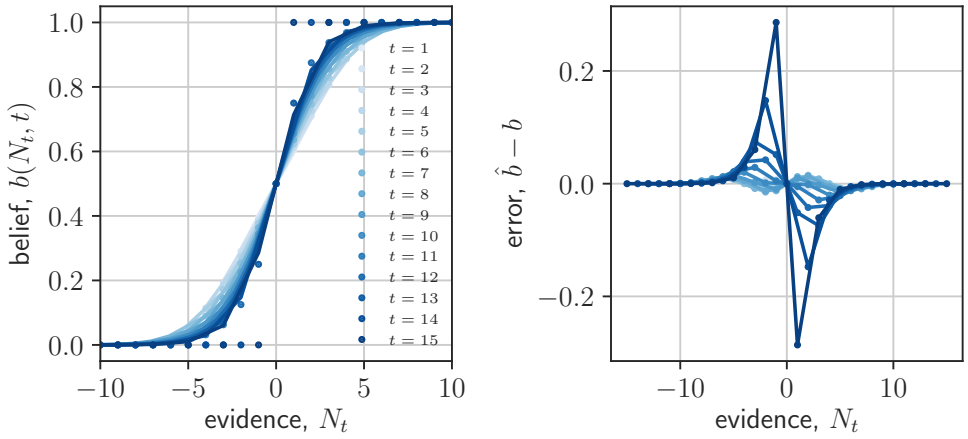

Figure S4. *sigmoidal approximation to expected reward.* (a) the approximation explained in Methods: State-conditioned expected trial reward, for different decision times. (b) The error in the approximation for different decision times.

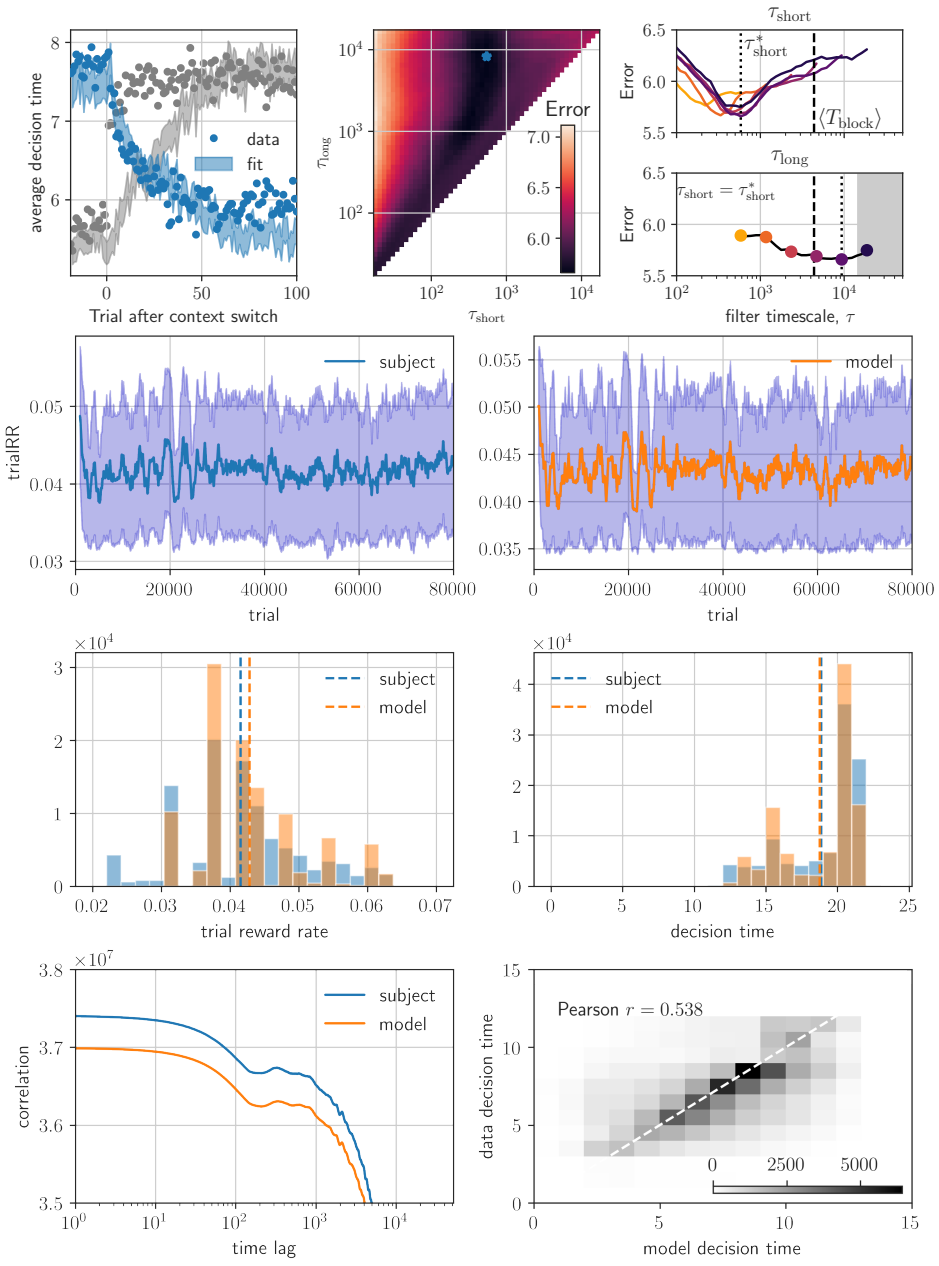

Figure S5. *Model validation on behavioural statistics from [S18].* Top: Running average trial reward rate $\rho_{\text{trial},k}$ over 1000 last trials. Middle: distributions of trial reward rate (left) and decision time (right). Bottom: Auto-correlation functions (left) and cross-correlation (right: gray-scale is trial count; white dashed line is perfect correlation)

.

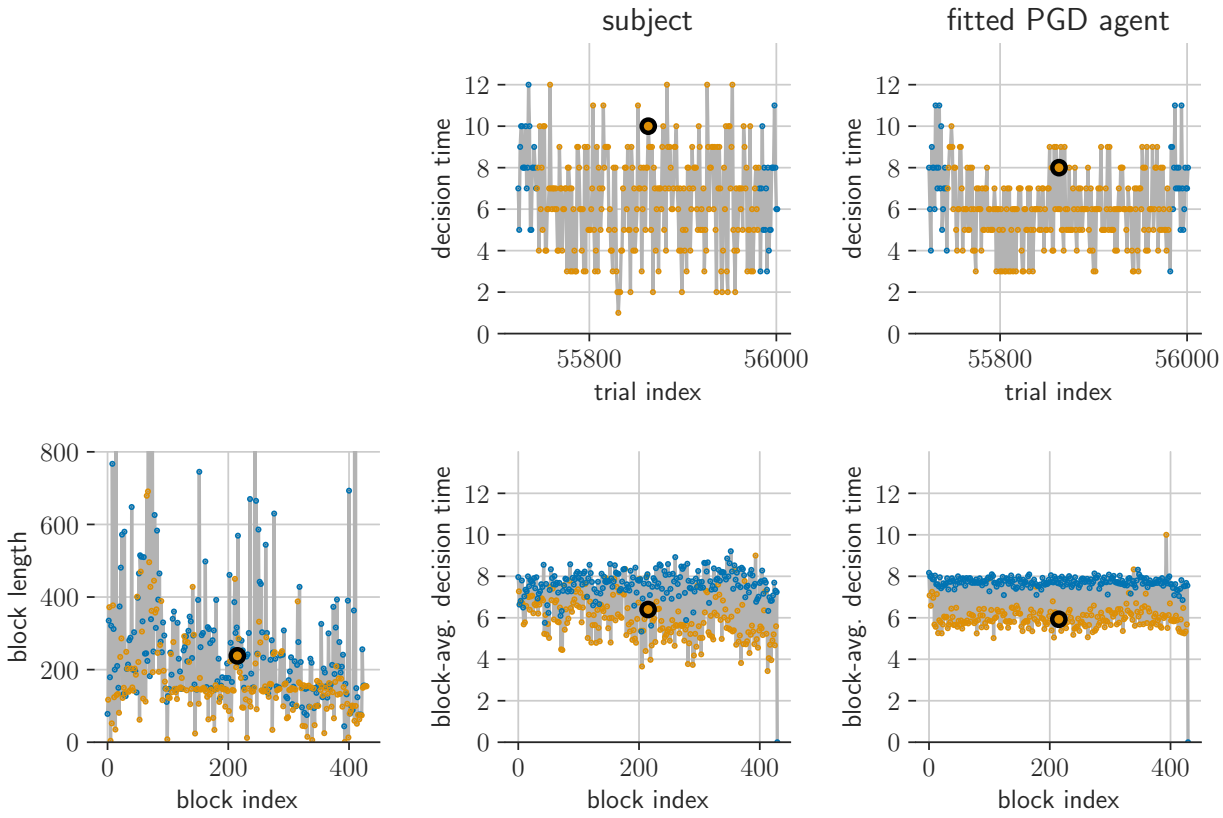

Figure S6. *Comparison of PGD and NHP in non-stationary α dynamics from [S18].* (a) The sequence of trial block durations used. (b,c) Decision times during a single block. (d,e) block-averaged decision times over the experiment.

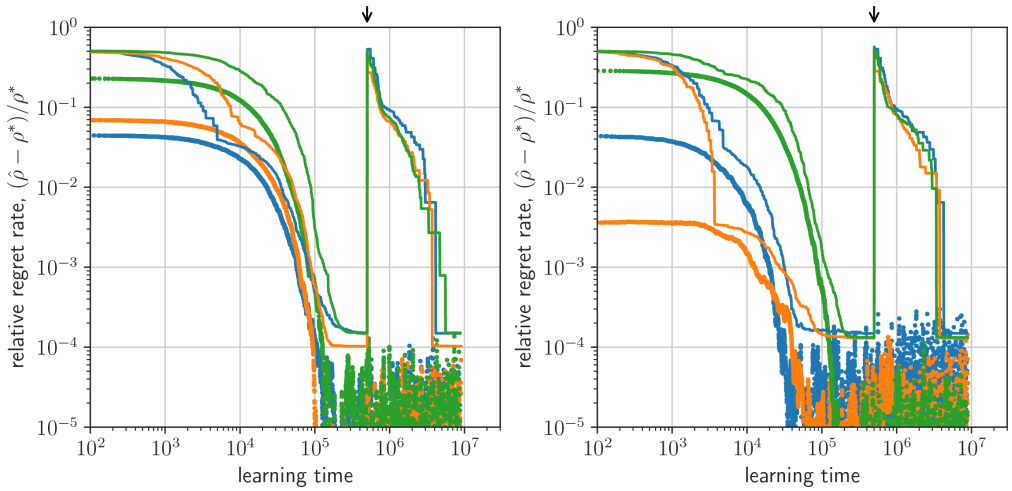

Figure S7. *Comparison of PGD and AR-RL learning on a patch leaving task.* Performance is defined as relative regret rate, $(\hat{\rho}-\rho^*)/\rho^*$ (PGD (dots); RL (lines)). (a) Performance over different sizes of the state vector ($d = 100$ (blue), 200 (orange), 300 (green)). (b) Performance over different learning rates (parametrized by integration time constant, $\tau = 1 \times 10^4$ (blue), 2×10^4 (orange), 3×10^4 (green)). (c) Schematic showing how to get from the stationary opportunity cost (the estimated reward rate, $\hat{\rho}_k^{\text{long}}$), to the decision boundary, $\boldsymbol{b}_t$. The PGD algorithm uses the opportunity cost directly, while value function methods require concurrently estimating a value function. (parameters: $\lambda = 1/5$; $r_{\max}$ sampled uniformly on $[0, 1]$). A random state label permutation is made at the time indicated by the black arrow.

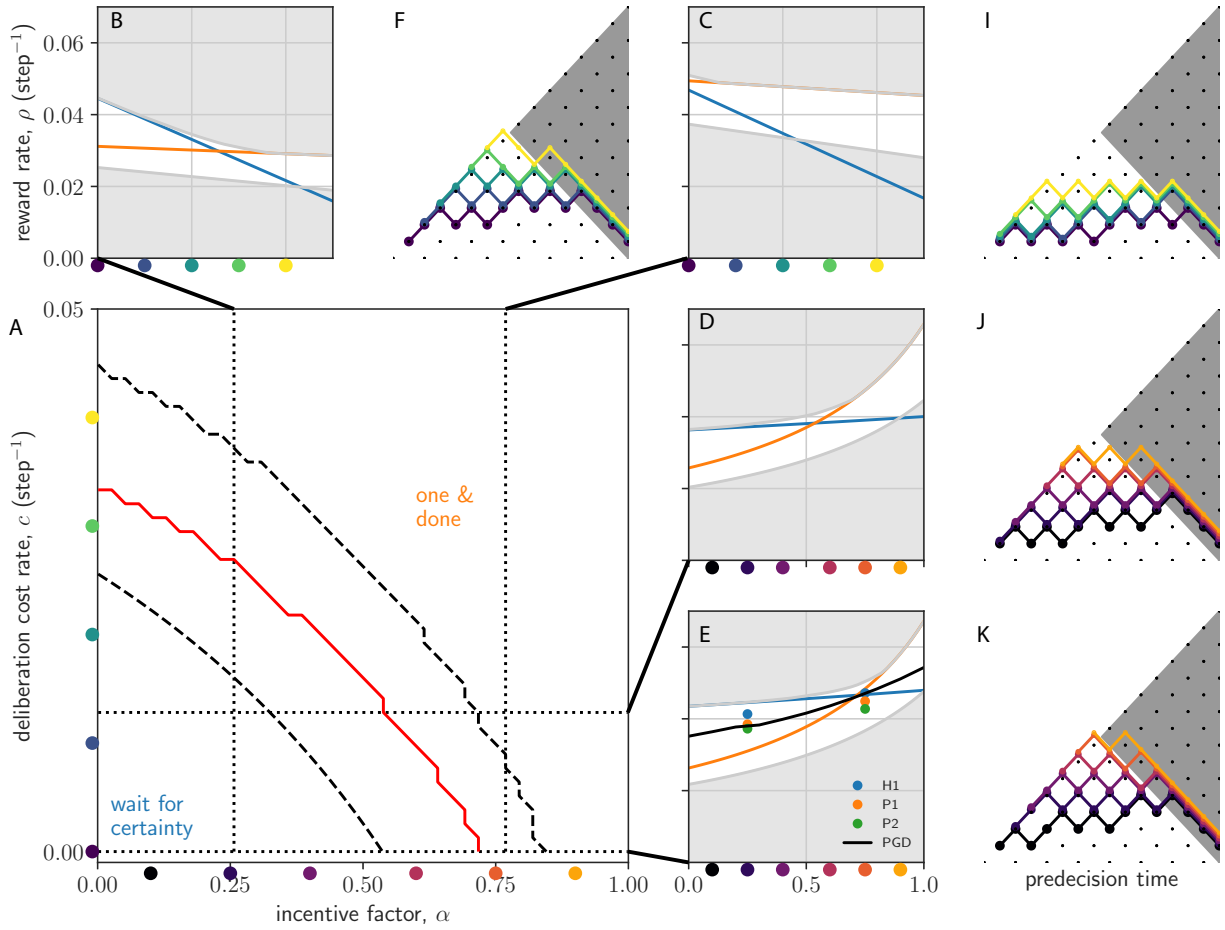

Figure S8. *Observed behaviour not in space of constant deliberation cost, reward rate-maximizing strategies.* (a) The reward-rate maximizing policy interpolates from the wait-for-certainty strategy at weak incentive (low α) and low deliberation cost (low c), to the one-and-done strategy at strong incentive (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 c (b,c) and α (d,e) (wait-for-certainty shown in blue; one-and-done shown in orange). In (e), we additionally show the context-conditioned reward rates for the two primates (P1,P2) as well as a reference human (H1), and the PGD algorithm (black line). Reward rates for primates are squarely in between the best and uniformly random strategy (lines bounding the upper and lower gray regions, respectively). Given the high overlap in the strategies (c.f. fig. 4f-k), the PGD algorithm performs similarly as the data. Note that, unlike primate data, all optimal strategies give no intermediate decision times at ambiguous ($N_t \approx 0$) states.

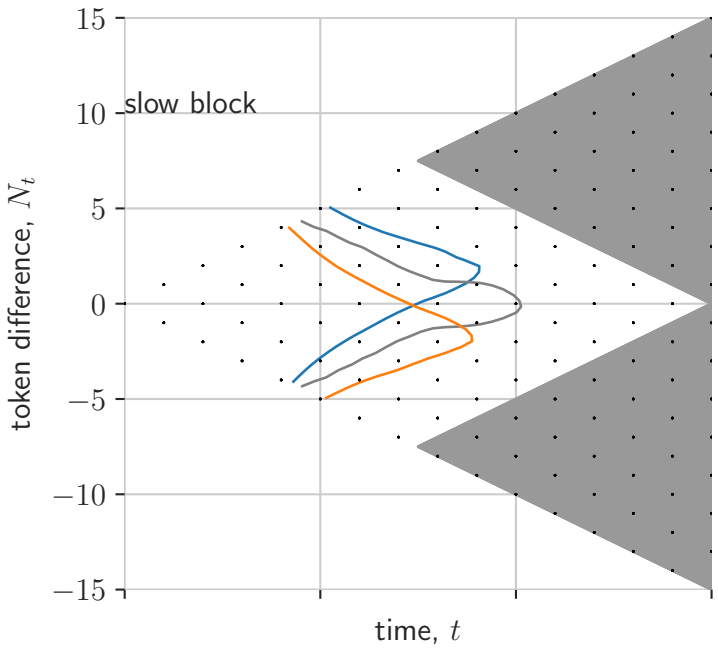

Figure S9. *Asymmetric action rewards skew survival probability.* Here, we plot the half-maximum of the PGD survival probability for three values of the action reward bias, $\gamma = -0.6, 0, 0.6$ (blue, black and orange, respectively). Other model parameters same as in fitted model.