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 obsessively attend to all details. Value it too much, and we carelessly rush to move on. How to strike this balance at each instant depends on what we know about our short and long term options and how reliable we think this knowledge is going forward. This challenging decision-making problem and its neural correlates are currently unknown. Here, we provide (1) a framework for this setting in terms of decision regret and opportunity cost; (2) a principled opportunity cost over decision hierarchies, which generalizes average-reward reinforcement learning; (3) a simple adaptive heuristic that handles non-stationarity by adaptively gating time investment according to this dynamic opportunity cost; and (4) a proposal—well-grounded in both cognitive and systems decision-making neuroscience—for how this heuristic is implemented in primate nervous systems. The proposal identifies urgency as the correlate of opportunity cost. 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. In particular, whenever one engages in a given activity that has some value, one necessarily foregoes other possible activities, which will also have some value. This is captured by the economic concept of "opportunity cost" [1]: the value of the alternative activities lost by committing one's time to any given activity. That cost is itself dependent on environmental properties that can change over time. For example, animals will learn to value a given food resource differently depending on whether it is encountered during times of plenty versus time of scarcity, producing behavior that may appear irrational when considered only within a single narrow context [2]. Incorporating opportunity costs into formal decision-making leads to a relative definition of reward. Average-reward reinforcement learning (AR-RL) is a reinforcement learning theory framework that uses the average reward as the reference [3]. It has been used to explain human and animal behaviour in foraging [4], free-operant conditioning [5], perceptual decision-making [6, 7], and cognitive effort/control [7, 8]. Up to now, this theory and its applications have been for stationary average reward, ie. for fixed context.

In psychological and neuroscientific studies of decision-making, we usually eliminate such contextual factors from the experimental design, presenting 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. 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, and which itself can be compared against a broader context of other things the subject could potentially do with their time. This view places the behaviour in the challenging continual learning setting at the forefront of contemporary artificial intelligence [9]. Importantly, the subject may not have full knowledge of these contextual factors and their stability or volatility. There are distinct rationales for how to handle this uncertainty depending on sensitivity to risk. For example, in situations where individual trials are short

and similar, one could simply estimate the average case and adjust behavior to that. At the other extreme, one could mitigate risk and always assume the worst-case scenario. Neuroscience is only beginning to address this average vs. worst-case distinction, the possible strategies of interpolating between the two, and the relevant neural correlates [10].

Here, we describe a simple heuristic strategy called Performance-gated deliberation (PGD), which trades off speed and accuracy on a given trial in a way that balances performance across a longer context that may be non-stationary and unknown. The algorithm separately learns 1) to estimate expected rewards on a given trial and 2) to estimate a context-aware opportunity cost that reflects performance across trials. This results in a simple "stopping rule" that defines when to cease deliberating and commit to a choice: the agent makes its decision when the opportunity cost of deliberation has exceeded the expected reward left to accrue. This heuristic can be interpreted as implementing a collapsing decision boundary probabilistic decision-making [6] and thus links to its putative neural correlate, "urgency" [6, 11–13]. These arise in policies designed around improving reward rate rather than more classical concepts of fixed accuracy criteria [14]. However, rather than feeding the opportunity cost into an optimization over a model of the task statistics as in AR-RL [6] or extracting urgency from data [15], PGD uses the opportunity cost directly as the decision boundary in a well-motivated approach 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. Briefly, the subject must guess 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, as soon as the subject makes their decision, the tokens accelerate to jump once every 50ms ("fast block") or once every 150ms ("slow block"), giving the subject the possibility to save time by taking an early guess. Behavior in the task, in both humans [12] and monkeys [15], is well explained by the "urgency gating model", which suggests that sensory evidence provided by the token jumps is combined with a growing context-dependent urgency signal to bias a competition between the two potential actions until one is selected as the winner. Neural recordings in monkeys suggest that evidence is estimated in dorsolateral prefrontal cortex [16], the urgency signal is provided by the basal ganglia [17], and the two are combined to bias a competition between potential actions that unfolds in dorsal premotor and primary motor cortex [15, 18]. Here, 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 balance between immediate rewards and total reward rates across multiple time scales.

RESULTS

Theory of performance-gated deliberation

Opportunity cost, regret, and an alternative to average-reward reinforcement learning

We focus on a class of continuing episodic tasks typically used in neuroscience. In this class, tasks consist of a long sequence of trials indexed by $k = 1, 2, \ldots$. In each trial, the subject gets to report once among a fixed set of options (see fig. 1a). Information about which option to choose for a reward is cumulatively accrued from evidence over the trial time t, and the subject must choose the moment to report its choice, denoted $t_{\mathrm{dec},k}$. Together with the chosen option, these determine both the reward it receives at the end of the trial, denoted R_k, and the duration of the trial, $T_k \geq t_{\mathrm{dec},k}$. Decision timing thus affects the rate at which the agent has the opportunity to accumulate reward. Such speed-accuracy trade-offs are central in determining performance in continuing task settings. For a fixed strategy, the *stationary reward rate* is then

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

Free-operant conditioning, patch leaving, and perceptual decision-making tasks often fall into this class. In the latter, the subject observes a sequence, $\boldsymbol{s}_t = (s_0, \ldots, s_t)$, of states, s_t, that provide evidence for a belief about the correct choice. The expected trial reward $\bar{r}(\boldsymbol{s}_t, t)$ over the residual uncertainty $p(r|\boldsymbol{s}_t, t_{\mathrm{dec}} = t)$ is conditioned on $\boldsymbol{s}_t$. Previous work [6, 19] has studied the choice probability $p(r|\boldsymbol{s}_t, t_{\mathrm{dec}} = t)$, but these are related: $\bar{r}(\boldsymbol{s}_t, t)$ is just the belief $p(r = 1|\boldsymbol{s}_t, t_{\mathrm{dec}} = t)$, since rewards are binary [6]. This reflects a broader connection between value-based and perceptual decisions [20]. We have suppressed conditioning these quantities on the choice taken since we will not explicitly address the problem of exploration and so assume the choice taken is always that with the largest belief at decision time.

For a given task, there is the maximum trial reward, $r_{\max}$, a subject can expect to receive a priori (assumed fixed here, e.g. a unit reward for a correct response). The *decision regret* at time t within a trial is given by the difference,

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

It decays (possibly non-monotonically) over time within a trial, as schematically shown in fig. 1b. $r_{\max}$ is fixed so minimizing $\mathcal{R}_t$ is equivalent to maximizing $\bar{r}(\boldsymbol{s}_t, t)$. However, the minimum possible regret is always 0, a property we will use below.

An agent lowers its regret towards zero here by waiting. Waiting, however, incurs opportunity cost: the reward lost by not acting. Previous neuroscience work on the opportunity cost of time [6, 7, 21, 22] have used the average-reward reinforcement learning (AR-RL) framework, first developed in artificial intelligence. Formally, the goal of AR-RL is to maximize the average-adjusted value $\mathbb{E}\left[\sum_{t'>t}(R_{t'} - \rho)\right]$, i.e. the expected sum of future reward *deviations* from the stationary average reward, ρ. To our knowledge, the subtraction of ρ was interpreted first in neuroscience as discounting reward by the opportunity cost rate [5], i.e. the expected reward forgone in each instant that the agent decides to wait. The accumulated opportunity cost in a trial up to a time t is then $\mathcal{O}_t = \rho t$.

For purposes of illustrating a simple example of AR-RL, take for a moment the specific setting where trials have no internal structure and are simply given by a fixed trial state, s,

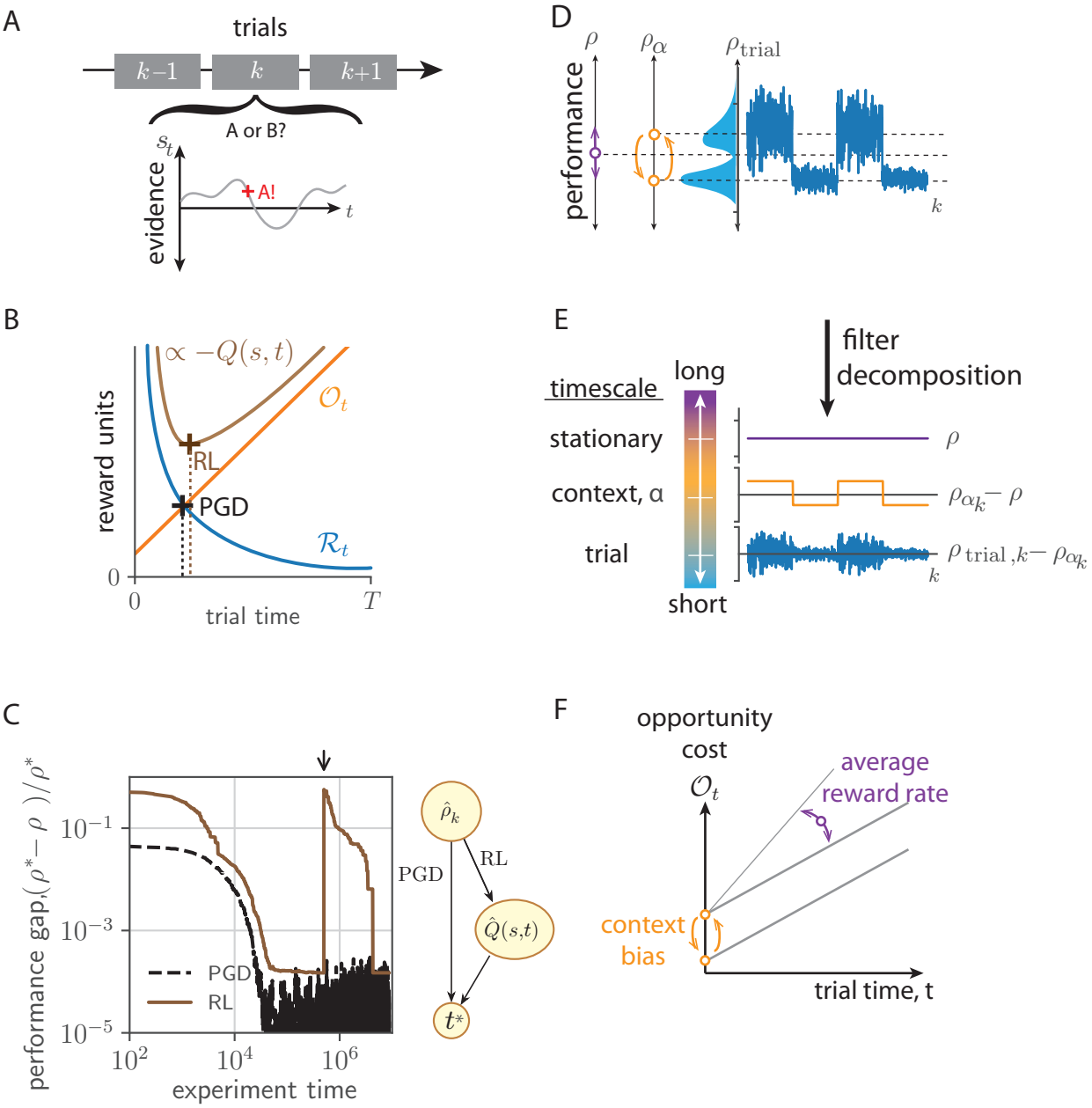

Figure 1. *Performance-gated deliberation.* (a) Task setting. Within trial evidence, s_t evolves over trial time t in successive trials indexed by k. A decision (e.g. 'A') is reported at the decision time. (b) Decision rules based on regret, $\mathcal{R}_t$ and opportunity cost, $\mathcal{O}_t$. The 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, $\mathcal{O}_t = \mathcal{R}_t$. (For simplicity, the task state is assumed fixed over t, $s_t = s$). (c) Convergence of performance towards optimum over learning in a patch foraging task. Shown is the performance gap , i.e. how much less is the 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: Algorithm schematic. Each uses the estimated reward rate $\hat{\rho}_k$. The AR-RL algorithm in addition estimates a value function. (d) Performance history (right) suggests context-dependent performance and exhibits multiple timescales. Contexts are evident in the reward distributions. (e) Trial, context, and effectively stationary timescales in an experiment form a hierarchy of components to reward history, decomposed by filtering. (f) The corresponding trial opportunity cost grows with slope ρ, and is offset by the context deviation, $(\rho_\alpha - \rho)T_\alpha$.

and the agent has the single choice of how long to set the trial, $T_k = t_{\text{dec},k}$. The action value $Q(s,t)$ then gives the average-adjusted trial return for deciding at time t in a trial of type s,

$$\begin{aligned} Q(s,t) =& \bar{r}(s,t) - \mathcal{O}_t \\ =& r_{\max} - (\mathcal{R}_t + \mathcal{O}_t) \ . \end{aligned} \tag{3}$$

This decomposition clearly shows the additive trade-off in jointly minimizing $\mathcal{O}_t$ and $\mathcal{R}_t$ in pursuit of maximizing value, $Q(s,t)$. The minimum of the sum thus occurs where $Q(s,t)$ achieves its maximum, giving the AR-RL optimal decision time. This novel perspective on the optimal average-adjusted value-based decision-making is shown in the schematic fig. 1b. From this perspective, an agent's solution to the speed-accuracy trade-off is thus given by how it balances decaying regret and growing opportunity cost.

A central feature of real world tasks is the possibility that task parameters might change. We argue that value representations, such as the average-adjusted value used in AR-RL (e.g. eq. (3)), are a liability in this non-stationary setting. This is because the value function has as many values as the number of states multiplied by the number of actions, which can be large. In RL algorithms aiming to model learning, $Q(s,a)$ is typically estimated online through experience, one correcting sample at a time, and so adapts slowly [23]. To better see the consequences of this slow adaptation, consider a continuing task where an animal feeds among a fixed set of diverse food patches, e.g. berry patches. The patch return is total berries consumed in patch s in a time t (analogous to $\bar{r}(s,t)$, but deterministic given s and directly observed). It saturates in time according to some given profile, as the fewer berries left are harder to find. The animal thus needs to decide when it is better off leaving the current partially depleted patch for another fully replenished patch (see Methods for task details). During its experience, the animal can estimate in the form of eq. (3), the value of being in a given patch for a given amount of time. However, if the patch environment changes (e.g. a plant disease randomly lowers yields), the estimated value $Q(s,t)$ will have to adapt and will do so slowly. We give an example of such slow adaptation with the brown line in fig. 1c, which shows how the performance gap narrows to the optimal reward rate ρ^* over learning. The performance of the value-based AR-RL algorithm improves down to the precision of its estimate for the reward rate ρ. However, a strong environment perturbation (here a random state label perturbation at the indicated time) leaves the AR-RL algorithm essentially back to where it started. It then has to relearn all state- and action-value associations. This drawback also afflicts alternative approaches that directly learn policies instead [24]. How then could high-value decision times be obtained without state-associations? We propose that instead of the maximum operation at the center of the AR-RL optimal solution, the agent simply take the intersection of $\mathcal{O}_t$ and $\mathcal{R}_t$ (shown as the black cross in fig. 1b). We call this algorithmic idea at the center of our results *performance-gated deliberation* (PGD). This means deciding as soon as opportunity cost exceeds decision regret. Plotted for this example patch leaving task in fig. 1c, PGD achieves better performance than AR-RL overall and is completely insensitive to such non-stationarity. 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 end this section by highlighting three aspects of this example that will come up again later in our main analysis of the tokens task.

First, 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 suboptimal at long times, 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, we argue this lower-level learning is typically stationary with respect to context variations in the way we will define them. PGD nevertheless retains its fast adaptation properties, making it an attractive heuristic strategy in environments with non-stationary context.

Second, coming back to neuroscience, we are particularly interested in studying PGD in deliberation tasks because a detailed picture of the underlying neural computations has been formed over decades of research [25, 26]. The PGD formulation easily extends to account for specific features of such tasks (e.g. dynamic state sequences and time penalties for incorrect responses [6]).

Third, 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. In this case, a broader notion of opportunity cost, one that accounts for multiple timescales, is needed. We present this setting in the following section. Using this dynamic opportunity cost in the PGD decision rule makes it applicable to a broad class of multiple timescale settings.

Filtering reward for a dynamic opportunity cost of time

Here we consider task parameters that vary across trials according to context and lead to distinct context-conditioned performance. We nevertheless assume that any such non-stationarity is at finite (though possible long) timescales, so that there is a well-defined time-averaged reward, ρ (e.g. seasonal effects are negligible when averaging over decades). For this use, we introduce a context parameter, α, whose value we assume has an effect on performance, i.e. the stationary reward for a fixed context, ρ_α, depends on α.

First consider α fixed. Then $\rho = \rho_\alpha$, $\mathcal{O}_t = \rho_\alpha t$, and the conventional average-reward reinforcement learning (AR-RL) case is recovered where ρ appears as a parameter that must be estimated. Previous works have used a low-pass filter to estimate reward rate [7, 27]. This filter simply averages over an exponentially-shaped window of past observations. For unit steps of discrete time, the step-wise update is

$$\hat{\rho}_t = (1-\beta)\hat{\rho}_{t-1} + \beta\rho_t \ , \tag{4}$$

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). The cumulative update of eq. (4) that smooths the reward uniformly over the trial duration and is applied once at the end of each trial reduces to [7, 27]

$$\hat{\rho}_k = (1-\beta)^{T_k}\hat{\rho}_{k-1} + (1-(1-\beta)^{T_k})\rho_{\text{trial},k} \text{ (online estimate)}, \tag{5}$$

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. (5) is used. This 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 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 $\hat{\rho}_k^{\tau_{\text{long}}}$.

Now consider a time-varying context as an α-sequence, α_k, with a single fixed, finite timescale, τ_{context}, over which it and thus the context typically remains the same. τ_{context} could refer to a deterministic periodicity or to rates of some Markov switching process, for example. τ_{context} can be learned adaptively from registering the frequency of change-points in the observed task structure, or even in performance if the change is not directly observable. This leads to a performance estimate, $\hat{\rho}_k^{\tau_{\text{context}}}$ that, unlike, $\hat{\rho}_k^{\tau_{\text{long}}}$, is intended to vary with time and tracks the effective instantaneous, context-specific performance ρ_{α_k}.

By forgetting quickly enough that its instantaneous bias is low (at the expense of its variance), $\hat{\rho}_k^{\tau_{\text{context}}}$ at first appears like straightforward way to extend the AR-RL formulation of stationary opportunity costs to the non-stationary case. However, if the agent knows the timescale of context dynamics, τ_{context}, that means the agent has the ability to plan on these timescales. Thus, a decision hierarchy arises from utilizing context knowledge, with moment-to-moment actions grouping into plans executed in particular contexts. But, the opportunity costs associated with these context plans are clearly not the same as those of moment-to-moment decisions, leaving this approach inconsistent. How then should opportunity costs in a decision hierarchy be formulated? To make progress on this subtle question, we go back to the foundations of the opportunity cost concept as the value of alternative actions. These are expectations using beliefs formed over what the agent considers reasonable to include as alternative actions. In Methods: Opportunity costs and relative value RL, we thus propose *Belief Average-Reward Reinforcement Learning* (BARRL) that replaces the average reward used as the reference in the average-adjusted value function of AR-RL with an expectation over the reward associated with alternative actions using beliefs given by the agent. Only when the agent incorporates the belief that the task is stationary, ie. that it lacks salient timescales beyond the moment-by-moment decisions does the belief average opportunity cost reduce to ρ, the stationary average-reward used in conventional AR-RL.

The practical innovation in BARRL relies on using a multiple timescale decomposition of performance as an effective decision hierarchy. 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. Each adds a zero time-average variation at the respective timescale to the sum of components below it (see fig. 1d). As with ρ, these finite timescale-resolved components of opportunity cost need to be estimated by the agent.

Applying this scheme to the context variation introduced above, the context-specific component is $(\rho_\alpha - \rho)T_\alpha$, with T_α the stationary average trial duration in the fixed context setting. The total opportunity cost shown in fig. 1f is

$$\mathcal{O}_t = \rho t + (\rho_\alpha - \rho)T_\alpha \text{ (context-aware opportunity cost)} , \tag{6}$$

where the second, novel term in eq. (6) is a baseline cost incurred at the beginning of each trial and computed as the deviation in opportunity cost accumulated over an trial on average from that context. This deviation fills the cost gap made by using the average reward rate in the moment-to-moment opportunity cost instead of the context-specific average reward. We can verify that this baseline has 0-mean using the mixed context ensemble average definition of ρ. In the case of large and roughly equal characteristic runs of trials from each context,

$$\rho \approx \sum_\alpha \rho_\alpha T_\alpha / \sum_\alpha T_\alpha \text{ (context-average)} , \tag{7}$$

from which $\sum_\alpha(\rho-\rho_\alpha)T_\alpha=0$ follows.

We propose $(\hat{\rho}_{k-1}^{\tau_{\text{context}}}-\hat{\rho}_{k-1}^{\tau_{\text{long}}})T_{k-1}$ to estimate eq. (6) where we have used the sample T_{k-1}, since the variance of T_k around T_α is small. We can then just time-integrate $\hat{\rho}_k^{\tau_{\text{context}}}-\hat{\rho}_k^{\tau_{\text{long}}}$ over the current trial. We refer the reader to fig. 6 for a mathematically precise filtering scheme that takes the reward sequence as input and outputs $\mathcal{O}_t$.

Neuroscience applications: PGD in the tokens task

We fapplied the PGD algorithm to the tokens task (see Methods: Tokens task description; fig. 2a), a random walk prediction task in which the incentive to decide early is controlled by a task parameter, α. In the experiments we analyze, α takes either a high or low value serving as a strong vs. weak incentive experimental condition (labelled the fast and slow context, respectively). Unlike the patch leaving task used above, here there are many within trial states and the state dynamics is stochastic. The expected decision regret (computed in Methods: State-conditioned expected trial reward) 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 (we come back to this feature in the discussion).

PGD in the stationary tokens task

We first show how the algorithm behaves when the dynamics of α is passive and simple. For this purpose, we consider an α sequence that switches back and forth between two values after a fixed number of trials, defining the trial block size (see fig. 2b,c). In this case, the agent model, i.e. its decision boundary and the quantities that determine it, relaxes into a self-consistent 2-block stationary noisy periodic trajectory (fig. 2g). We set τ_{context} at a few tens of trials and τ_{long} two orders of magnitude larger so that it averages over tens of blocks. The decision times relax after a context switch (fig. 2e) to their conditional average but exhibit strong fluctuations from the random sequence of random walk realizations. The switching is stationary however (fig. 2c), and so 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). Indeed, the resulting estimates ρ_{long} and ρ_{context}, are near their stationary values (dashed lines in fig. 2g,h). While these estimates improve 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 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). Matching ρ_{long} to ρ, at least to the precision possible, satisfies the algorithm's self-consistency that arises from the dependence of the decision boundary, which determines the performance, on the performance itself. Self-consistency also arises in AR-RL, where it is used to determine the average reward rate (see, e.g. [6]). On the other hand, if $T_{\text{block}}/\tau_{\text{long}}$ is large, ρ_{long} approaches ρ_{short} and opportunity cost is We propose this limit as a test of the theory in the discussion.

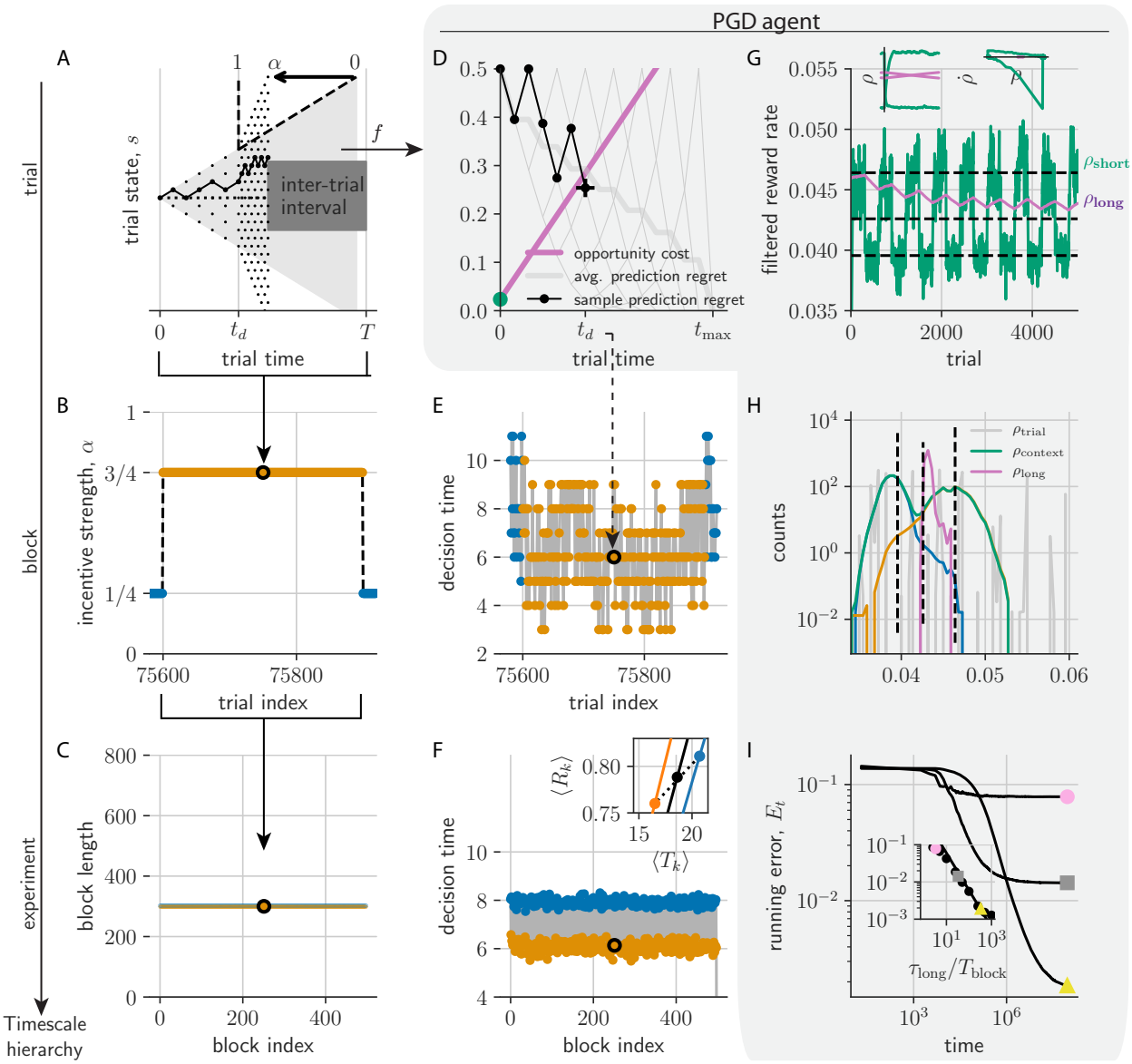

Figure 2. *PGD agent plays the tokens task with periodic α-dynamics.* (a) A tokens task trial for $\alpha = 3/4$ and decision time t_d. (b) Trials are grouped into alternating trial blocks of constant α (fast (orange) and slow (blue) conditions). (c) Here, trial block durations are constant over the 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). 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).

Next, we applied the PGD algorithm to the actual sequence of random walks and incentive strengths used in the experiments reported in [18] (see fig. 4a). As in the above example, trials alternated between a fast and slow context, but here with irregular block size determined by the discretion of the experimenter who accommodated for the animal's motivation and had target minimums for the block size [28]. The block size statistics for this animal were generally over-dispersed and had weak auto-correlation.

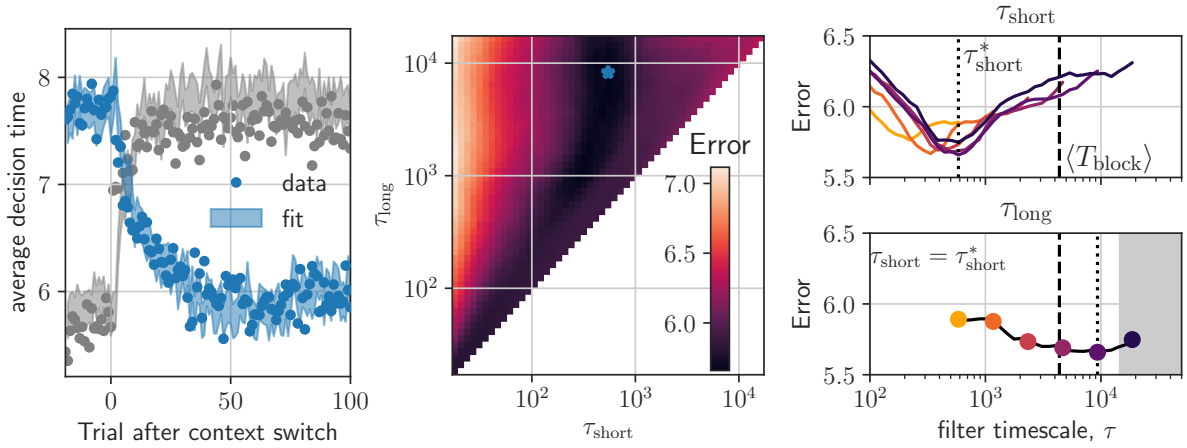

Figure 3. *Model fit.* (a) Average decision times (dots) aligned by context switch. Shaded region is standard error of model, which includes the asymmetric switching component in the Methods. (b) Error surface over τ_{context} and τ_{long}. Model fitted only to slow-fast transition (blue), not fast-slow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

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). We found a hysteresis (asymmetric relaxation timescales after context switches), with the fast-to-slow switch happening almost instantaneously. We thus focussed on the slow-to-fast transition to set τ_{context}. Nevertheless, we developed a simple model accounting for the asymmetric switching costs that captures the relaxations after both switch types with addition of a single sensitivity parameter (see Methods: Assymetric switching cost model). We note that the latter was not necessary to capture to fit τ_{context}. We found that these precisely identified τ_{context}, but they 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. 10).

If we condition the data set on α we get two context-specific strategies. While decision times are informative of strategy, the full strategy depends also on the environment state at decision time (see fig. 4). We can view the action policies computed from the histograms of $(N_{t_{\text{dec}}}, t_{\text{dec}})$, over trials. However, these histograms 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 mapping to the ensemble of state conditioned on time $(N_t|t)$ and the event that $t = t_{\text{dec}}$. Conditioning this ensemble on the state gives

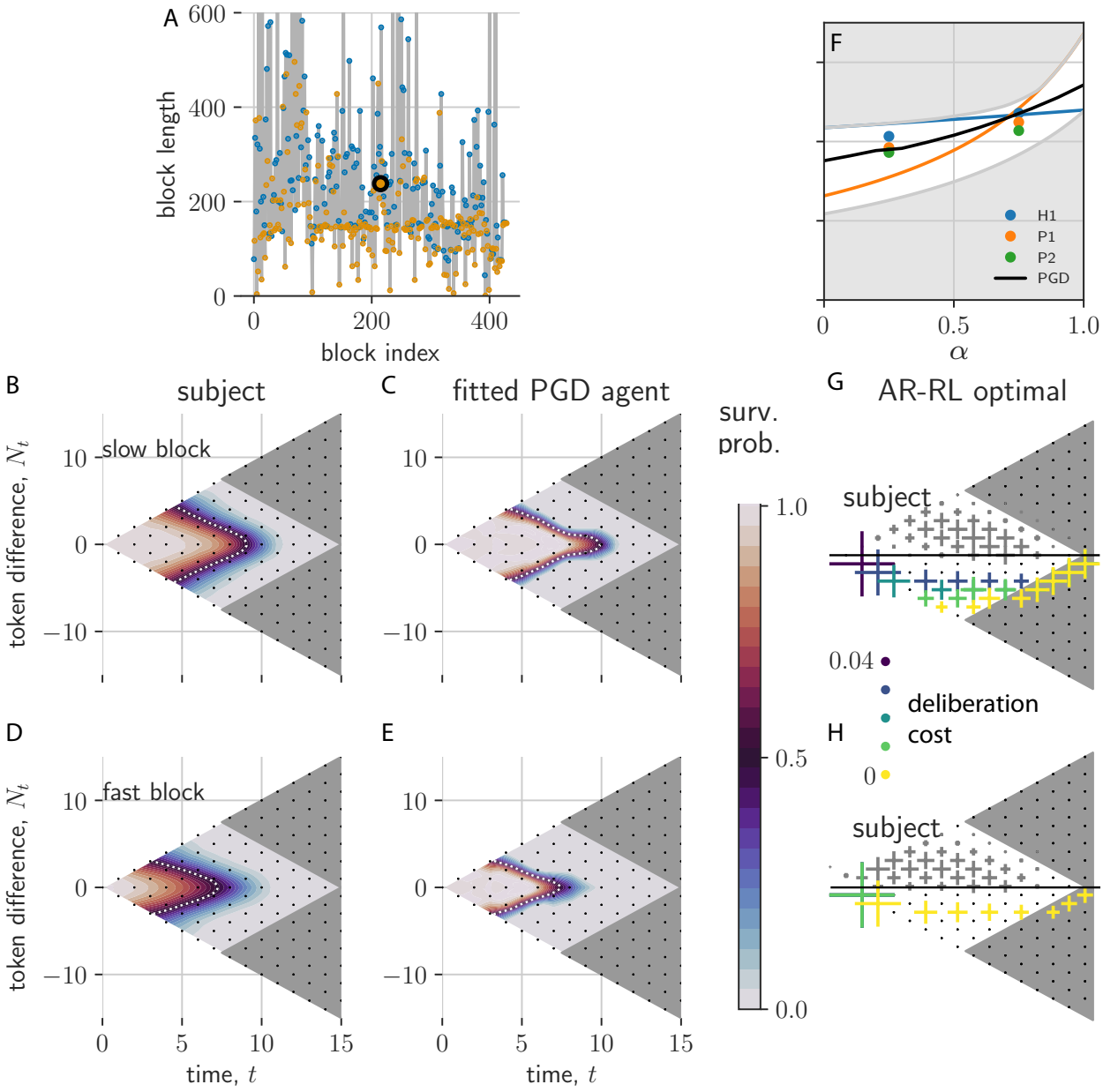

Figure 4. *Comparison of PGD and NHP in non-stationary α dynamics from [18].* (a) block length sequence used in experiment (c.f. fig. 2c). (b-e) Interpolated state-conditioned survival probabilities over slow (b,c) and fast (d,e) blocks. White dashed lines show the half-maximum ($P = 0.5$) contour. (f) Shown are the reward rate as a function of incentive strength and no deliberation cost ($c = 0$) (wait-for-certainty shown in blue; one&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) Decision time 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..

$P(t = t_{\text{dec}}|N_t, t) = p(N_t, t = t_{\text{dec}}|t)/p(N_t|t)$. To reduce estimator variance, we focus instead 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, the result, fig. 4b-e shows a remarkably smooth average strategy, shared between both the model and the animal behaviour. Fast block strategies are shifted earlier by similar amounts relative to slow block strategies in both model and data. Given the almost single degree of freedom in the model (τ_{context}), the strong, almost quantitative correspondence is encouraging. Any quantitative approach requires situating the algorithm in the behaving animal by adding features such as motivational noise. We leave this to future work.

To better understand where both the data and the PGD agent lie in the space of strategies, we computed reward-rate optimal solutions for stationary contexts using average-adjusted value functions (Methods: Episodic decision-making and dynamic programming solutions of value iteration). We also account for a constant deliberation cost rate, c [6]. Over the (α, c)-plane, the optimal solution interpolates from the wait-for-certainty strategy at low α and c to the one&done strategy [29] at high α and c (see fig. 13). Importantly, the family of strategies across c for the slow $\alpha = 0.25$ and fast $\alpha = 0.75$ condition (fig. 4g,h) are qualitatively distinct from the data and the PGD strategies. Given the high overlap in the strategies (b-e), the PGD algorithm performs similarly as the data. Indeed over all α and c, the strategies near the end of the trial wait to resolve the remaining ambiguity. This is in contrast to the primate and PGD behaviour that resolve residual ambiguity at intermediate trial times (see e.g. fig. 4b-e). This performance falls in between this optimal and the random MDP strategy that picks one of the three actions (report left, report right, and wait) at random (see fig. 13f). Given the good model fit, these properties of the behaviour are of course shared by the PGD agent, whose performance sits in between the random and optimal performance across the range of values of α.

Neural urgency and opportunity cost

Here, we take up the important step of grounding the above theory of behaviour in the dynamics of the brains that produce it. This serves as a means to test our prediction that neural urgency reflects our formulation of a dynamic opportunity cost. In fig. 5a, we restate in a schematic diagram the consensus understanding of the neural dynamics of decision-making [30]. In fig. 5b, we show the decision-making area (here PMd)'s population firing rate conditioned on zero-evidence environment states during trials from the data in [18]. With an affine unit conversion from reward to spikes/step (here simply distinct y-axes), the conditioned opportunity cost signals map tightly onto the observed urgency signals. This correspondence has multiple features: (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. Namely, that (1) there is a constant step-wise opportunity cost rate, (2) there is a opportunity cost incurred over larger times, and (3) this cost varies proportionally with the context-specific reward rate.

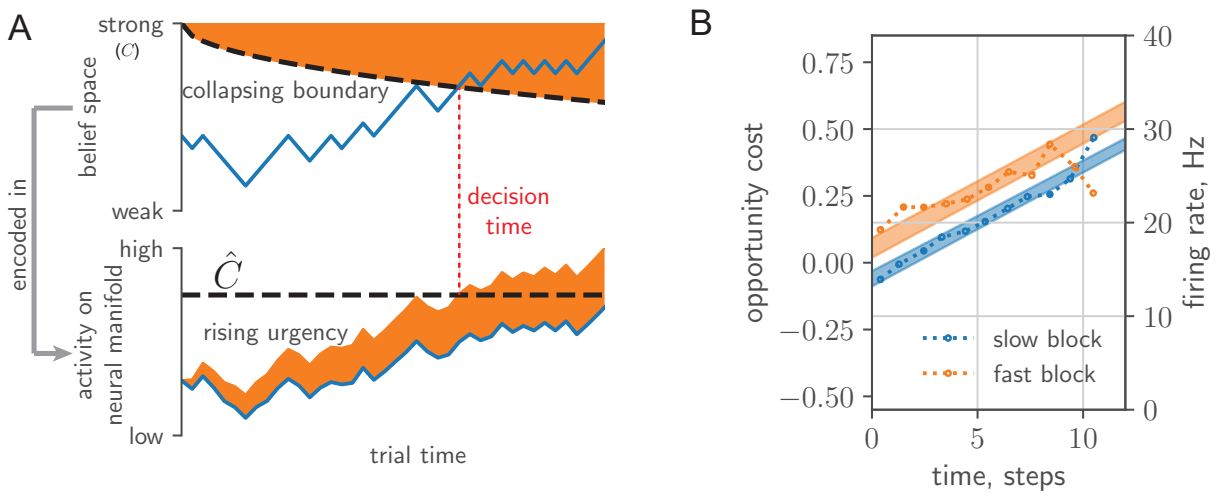

Figure 5. *Neural urgency and collapsing decision boundaries and comparison with data.* (a) Top: evidence accumulation in belief space. Decisions are made when the strength of the belief, E_t, exceeds a desired confidence. This confidence is lowered from a (possibly time-dependent) baseline confidence (here fixed at C) according to some function, $-U_t$. Bottom: Belief is encoded into a low-dimensional manifold (e.g. in 1D as log odds; we denote $\hat{x}$ as the neural encoding of x). Manifold dynamics integrates neural evidence as well as a rate of urgency such that in a simple instantiation, the two time-integrated signals combine linearly, $\hat{E}_t + \hat{U}_t$. Once manifold activity exceeds some fixed value $\hat{C}$, a decision is effected via the projections into downstream motor areas, which begin execution of the associated motor response. (c.f. fig.8 in [6].) (b) Opportunity cost linearly maps onto the urgency signal extracted from 0-evidence conditioned cell-averaged firing rate in PMd (c.f. fig.2f [18]).

DISCUSSION

Here, we have proposed a heuristic decision-making algorithm that gates deliberation based on performance. While we gave a foraging task example in which it is optimal in the average-reward reinforcement learning (AR-RL) sense, it is generally not optimal in either of the average or worst case sense. It is however generally applicable and at once exploits the stationarity of the environment statistics and hedges against non-stationarity in the incentive strength, striking a balance between strategy complexity and return. For example, by splitting the problem into two separate components–learning the statistics of the environment and tracking one's own performance in that environment–only parts of the algorithm need adjustment upon task structure variation (reminiscent of the decoupling of complex state dynamics from reward achieved by successor representations [31]). The result is a robust heuristic that at once balances time investment and handles performance-relevant non-stationary task variation.

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. We used behavioural data to shape the theory, and neural recordings to provide evidence of one the neural correlates it proposes: the temporal profile of neural urgency in decision-making areas.

The broader proposal is of a cortico-basal ganglia system in which performance on mul-

tiple, behaviour-relevant timescales are broadcast to multiple decision-making areas to gate decision speed. While the view that tonic dopamine encodes average reward is decades old, the multiple timescale representation has received substantial empirical support in recent years, from cognitive results [32–34] to a recent unified view of dopamine signalling of temporal difference errors using multiple discount factors [35] and of dopamine as encoding both value and uncertainty [36]. While the relationship of dopamine to time perception, and thus putatively to decision speed has been proposed [37] in the reward learning literature, a proposal that includes how speed is implicated in the decision-making areas driving the actions has been absent. Through (at the present time) unspecified projection and/or gating mechanisms that suitably combine the estimates from distinct timescales, our proposal is that decision-making areas across the brain are driven by this multi-channel performance signal to gate the speed of their attractor-based decision-making dynamics [38, 39]. The lack of value learning in PGD is also consistent with recent work suggesting it is often unnecessary [24]. In this sense, our theory goes beyond existing proposals for the role of dopamine signalling in coding for reward rate by considering dynamic evidence tasks and mechanistically identifying urgency as the means by which the reward code ultimately affects decisions. The theory also has clinical implications as it ties beliefs (about rewards) to vigor (speed of decision-making) and thus identifies opportunity cost estimation as the causal link between the observed correlation between vigor impairments (e.g. in Parkinson's disease) and dysregulated dopamine signalling in the reward system [40–42]. Furthermore, psychological test batteries are demonstrating the role of urgency as a transdiagnostic indicator of cognitive impairments related to impulsivity. Our theory offers a mechanistic means to ground these results in neural dynamics and, crucially, how it is impaired.

We gave concrete proposals for the means to learn the opportunity cost, and simply 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 [16]. For the general class of tasks we consider, a generic, neurally plausible means to learn $p_{n,t}$ is via distributional value codes (see e.g. [36]). The Laplace code is a distributional value representation (implementable via an ensemble of successor representations) that uses an ensemble of units over a range of temporal discount factors and reward sensitivities[43]. The expected reward at a chosen future time can be easily computed using this representation. We are pursuing this approach in a follow-up study.

The theory is prescriptive at both the behavioural and neural level with regards to the shape of the action policy (e.g. fig. 4) and temporal profile of neural urgency. The survival probability representation we devised is a robust measure of mean performance in tasks where the evidence available to the animal is known or can be inferred. It also varies markedly with reward structure, and thus provides a wealth of predictions. At the behavioural level, for example, a salient feature of the behavioural policies studied here are their reflective symmetry in tokens difference, N_t. From the perspective of the theory, this arises from the symmetry in the payoff matrix of rewards. We can break this symmetry in the theory for which the resulting behavioural strategies take on a distinctly asymmetric shape (See Methods: Asymmetric rewards). At the neural level, our theory prescribes the temporal profile of the urgency signal. As with behaviour, the theory makes clear predictions about how it should vary with changing task structure. A simple prediction is that for fixed parameters, the baseline opportunity cost shrinks while the variation of the slope should grow with increasing the length of the trial block, e.g. for block sizes larger than

$T_{\text{block}} \approx (\rho_\alpha - \rho)\tau_{\text{long}}$ at which $\hat{\rho}_k^{\tau_{\text{long}}} \approx \hat{\rho}_k^{\tau_{\text{context}}}$ most of the time. At transitions between these long blocks, the baseline component is phasic, while the slope relaxes monotonically to its stationary value for that block type. This prediction presumes the animal doesn't adapt too much so that the PGD parameters can be assumed fixed. The data we analyzed was for block lengths short enough that the slope remain fixed across blocks. The concrete prediction is that much longer blocks should start to exhibit variation in slope between block type.

The decision boundaries generated by PGD have the same qualitative dependence on context as those of the reward rate optimal strategy, while in addition adapting much faster to context switches than typical online versions like TD-learning. Unlike worst-case optimal algorithms that eschew environmental information, our algorithm exploits the stationary behaviour of the trial environment such that its decision boundaries reflect precise information about reward statistics. For all these benefits, we believe that relatively simple and nimble strategies such as the one we propose 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.

Our work impacts modern reinforcement learning, by generalizing average reward RL to a belief average. We have left a detailed algorithmic analysis of BARRL to future work, but expect improvements in similar settings as successor representations. Indeed, our work suggests a new class of algorithm between model-based and model-free. Moreover, the average-reward framework is in many ways more suitable for the continuing task setting than the discounted approach with which the conventional successor representation is defined.

It remains to explain when PGD is likely to be used by an agent. PGD fills a intermediate space between exploiting detailed knowledge and using simple heuristics. Highly incentivized human behaviour is likely more structured because of access to more sophisticated learning. For example, offline learning such as experience replay buffers can greatly improve performance in machine algorithms and this also probably true in animals. There also is the possibility that PGD, despite its bias, is to some degree a hard-coded heuristic strategy. Hard-coded or not, the question remains if it is optimal with respect to some bounded rational objective, an interesting direction for future work. We raised the problems with this approach in the introduction. Nevertheless, the question remains that humans, despite our apparent access to sophisticated computation, exhibit measurable bias in how we incorporate past experience [44]. There is an interesting evolutionary perspective on the individual bias that adapting strategies, like PGD, exhibit. One simple example is the win-stay/lose-shift strategy, a more rudimentary kind of performance-gated decision-making than PGD. This strategy was shown to explain how humans approach the rock-paper-scissors game [45]. 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 knowledge-poor strategies can have surprisingly good evolutionary benefit.

leftover points for discussion:

- The places where ad hoc modelling decisions had to be made and alternative choices would change the result.

- stochastic versus deterministic policies

METHODS

Tokens task: description and properties

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 $t_{\max} - t_{\text{dec}}$ jumps after t_{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

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

where a dead time between episodes, T_{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. 13j,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. 9. 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.

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

$$\rho = \sum_s^d r_s(t_s) \Big/ \sum_s^d t_s \text{ (patch average)}. \tag{19}$$

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. (19) 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. 12. 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 RL algorithm. Approaching this bound, both PGD and RL learning convergence time is limited the integration time τ of the estimate $\hat{\rho}_k^\tau$ (c.f. eq. (5)) 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 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.

Opportunity costs and relative value-based reinforcement learning

In this section, we operationalize opportunity cost from first principles within a reinforcement learning setting leading to Belief Average-Reward Reinforcement Learning (BAR-RL), which uses the well-known framework of Average-Reward Reinforcement Learning (AR-RL) [3] and reduces to it in the case of stationary tasks and flat decision hierarchies. In economics, opportunity costs are typically defined as the value forfeited when committing a given resource to a specific use. In this setting, the resource is the action sequence an agent takes in the world, and the commitment at each time is to the selected action. What is given up at each time step is then the rewards received from having instead taken some alternative

action. Costs are typically formulated as negative rewards. However, reinforcement learning theory was formulated to capture the delayed rewards received at future times that are associated with taking an action when in a given state. It does so by defining the *value* of an action via the expected sum of future rewards, rather than just the immediate reward. It is this value of an alternative action that is forfeited when selecting another, and not just the alternative action's immediate reward. The value of the alternative action is then what constitutes the opportunity cost of the selected action in the reinforcement learning setting.

A straightforward relative value definition employs the same value definition for each of the pair of compared values. Our formulation deals directly with the most direct definition of *state-action value*,

$$Q(s,a) = \mathbb{E}^{\pi}\left[\sum_{k=1}^{\infty} R_{t+k}\middle| S_t = s, A_t = a\right] . \tag{20}$$

Here, we have modelled the agent-environment system as a discrete-time Markov decision process. The state and action sequence, S_t and A_t, are part of this process for which $\mathbb{E}^{\pi}$ applies the expectation over the k-indexed future state sequence, S_{t+k}, observed when following an action policy π after choosing (i.e. conditioned on) action a in state s. The time index of this event is labelled t, and its immediate reward (possibly zero) is $R_{t+1} = R(s_t, a_t)$, where $R(s,a)$ is the reward function. The value expressed in eq. (20) is in general unbounded because of the infinite sum. In practise, standard reinforcement learning employs a discount-adjusted value, a modified version of eq. (20) in which the kth term is multiplied by γ^{k-1}, where γ is a free parameter called the discount factor. This imposes an *ad hoc* effective horizon time, $1/(1-\gamma)$, a time into the future within which rewards are considered and beyond which rewards are ignored. For a general class of tasks, however, the statistics of R_{t+k} relax to their unconditioned values with k as the memory of this event at t fades. Thus, when k is large, $\mathbb{E}^{\pi}\left[R_{t+k}\middle| S_t = s, A_t = a\right] \approx \mathbb{E}^{\pi}\left[R_{t+k}\right] = \rho$ where ρ is the average reward obtained when choosing actions according to π. This suggests a mean-subtracted formulation that naturally bounds the sum without the need for an additional free parameter such as the discount factor.

The *relative value*, denoted with a tilde, $\tilde{Q}(s,a)$, is simply the state-action value relative to a reference value, $O(s,a)$,

$$\tilde{Q}(s,a) = Q(s,a) - O(s,a) . \tag{21}$$

For example, when O is the true opportunity cost, i.e. the highest value among alternative actions, we have

$$\begin{aligned} O(s,a) :=& \max_{a' \neq a} Q(s,a') \\ =& \mathbb{E}^{\pi}\left[\sum_{k=1}^{\infty} R_{t+k}\middle| S_t = s, A_t = \operatorname*{argmax}_{a' \neq a} Q(s,a')\right] . \end{aligned} \tag{22}$$

Even though $Q(s,a)$ and $O(s,a)$ diverge in general, they appear in our formulation always in a difference, which is finite because of fading memory: for large enough k, the effect of which action is selected at time t is negligible. So for large k, R_{t+k} is distributed in both eq. (20) and eq. (22) according to the unconditioned stationary statistics of the process

and so, pulling out the sum in eq. (21), the difference of the pair of k-indexed expectations vanish. The transient behaviour of the two sums is thus what determines the magnitude of $\tilde{Q}(s,a)$. This motivates recasting reward-like sequences as deviations from ρ. Let us from hereon suppress denoting the conditioning on a and s and denote the random opportunity cost sequence (with its the conditioning the alternative action chosen in eq. (22)) as ρ_t. We then write $R_t = \rho + \delta R_t$ and $\rho_t = \rho + \delta\rho_t$. The state-action value becomes,

$$\begin{aligned}\tilde{Q}(s,a) =& \mathbb{E}^\pi\left[\sum_{k=1}^{\infty} R_{t+k} - \rho_{t+k}\right] \\ =& \mathbb{E}^\pi\left[\sum_{k=1}^{\infty} \delta R_{t+k} - \delta\rho_{t+k}\right] ,\end{aligned} \tag{23}$$

where in general $\mathbb{E}^\pi\left[\delta R_{t+k}\right] \to 0$ and $\mathbb{E}^\pi\left[\delta\rho_{t+k}\right] \to 0$ with k and their difference does so fast enough that $\tilde{Q}(s,a)$ remains finite.

We now focus on using eq. (23) to motivate alternatives and approximations to eq. (22) that arise from differing degrees of knowledge about rewards. Equation (23) immediately suggests the approximation $\delta\rho_{t+k} = 0$ for all k, which recovers the average-adjusted value function on which AR-RL is based. In light of eq. (22), this case results from replacing the maximum operation with an average over the state using the stationary state distribution arising from action policy π so that in effect no conditioning on state or action is applied when taking the expectation. The latter is equivalent to the Bayesian prior belief average when the state is considered unobserved and so this case is the minimal information choice for opportunity cost, highlighting a novel perspective on AR-RL. This choice simply places a fixed cost on time. Rewards are then incorporated into value via their magnitude relative to how valuable is a unit of time. This opportunity cost of time depends on the policy and is given by the average reward, ρ.

Conversely, when finite state information is made available, $\delta\rho_t$ is not always zero. As a result, the cost of time varies in time. This case can be understood by viewing $\delta\rho_t$ as a dynamic perturbation to the AR-RL case above since an equivalent form of eq. (23) is

$$\tilde{Q}(s,a) = \mathbb{E}^\pi\left[\sum_{k=1}^{\infty} R_{t+k} - (\rho + \delta\rho_{t+k})\right] . \tag{24}$$

Indeed, we can still use the average-reward formulation to analyze these extensions by absorbing $\delta\rho_t$ into an augmented reward sequence, $R'_t = R_t - \delta\rho_t$.

Unlike the discounted-reward formulation, our use here of the AR-RL formulation allows for solving the problem by considering only a single episode. Indeed, by leveraging AR-RL, our formulation of relative-value decision-making offers the multiple advantages over the conventional discounted-reward formulation in addition to this advantage for episodic tasks. First, the decay timescale of $\tilde{Q}$ is intrinsic to the agent-environment dynamics, in contrast to discounted reward formulations that achieve finite value by imposing an arbitrary finite horizon (explicitly via a horizon time or effectively via a discount factor). Finite horizons are nevertheless practical, and our relative value formulation can easily include them by setting them after the intrinsic decay timescale so they have no affect on the formulation. Second, the average reward, ρ, is explicit and so can be directly optimized (achieving so-called *gain optimal* policies), rather than approximated in the discounted reward formulation by choosing the discount factor close enough to 1. Third, optimizing the relative value function

explicitly optimizes transient behaviour (socalled *bias optimal* policies), in addition to ρ. Achieving this in the discounted reward formulation is difficult because numerical precision is lost when using discount factors near 1 and the effective horizon is large (the brittle numerics and slow convergence of the Neumann series underlying the discount-adjusted approach is well-appreciated [46]).

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 [6], 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,

$$Q(s, a|t) = \mathbb{E}^{\pi}\left[\sum_{t'=t}^{T} R_t - \rho \,\middle|\, S_t = s, A_t = a\right] + V(s|t = T + 1) \tag{25}$$

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. (25) 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

$$\begin{aligned} V(s|t) &= \max_a Q(s,a|t) \ , \\ Q(s,\pm|t) &= R(s,\pm|t) - C(t) + V(s|t=T+1) \ , \\ Q(s,wait|t) &= -c(t) + \mathbb{E}_{s_{t+1}|s}\left[V(s_{t+1}|t+1)\right] \ , \\ V(s|t=0) &= V(s|t=T+1) \ . \end{aligned} \tag{26}$$

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. (26) 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. (26) [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 (the case of $\delta\rho_t = 0$ considered in the previous section), $c(t) = \rho$. 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, $V(s|t=0)$, is equal to $V(s|t=T+1)$. This dependence can be seen formally by taking the state-action value eq. (25), choosing a according to π to obtain the state value, $V(s|t)$, and evaluating it for $t = 0$,

$$V(s|t=0) = \mathbb{E}_{t_d}\left[\sum_{t=0}^{T} R_t - \rho\right] + V(s|t=T+1) \tag{27}$$

$$= \mathbb{E}_{t_d}\left[R(t_d) - \rho T(t_d)\right] + V(s|t=T+1) \tag{28}$$

$$= \bar{R} - \rho\bar{T} + V(s|t=T+1) \ . \tag{29}$$

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. (29) gives $\rho = \bar{R}/\bar{T}$.

The expected trial return at decision time is the argument of the trial-average in eq. (28), $R(t_d) - \rho T(t_d)$, 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 average-reward 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 dynamic opportunity cost rates, i.e. varying cost place on time, by replacing ρ by its time-varying value, $\rho + \delta\rho_t$ (c.f. eq. (24)). It the trials are a independent and identically distributed sequence, 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. For example, if trials are sampled from

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. 8).

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

$$\tau_{\text{context}} = \frac{\tau_{\text{ref}}}{1 + C^{\nu}} \ , \tag{30}$$

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

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

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. 14. 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] Esteban Freidin and Alex Kacelnik, "Rational Choice, Context Dependence, and the Value of Information in European Starlings (Sturnus vulgaris)," Science **334**, 1000 LP – 1002 (2011).

[3] 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].

[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, Anne K Churchland, Michael N Shadlen, and Alexandre Pouget, "The Cost of Accumulating Evidence in Perceptual Decision Making," **32**, 3612–3628 (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] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup, "Towards Continual Reinforcement Learning: A Review and Perspectives," (2020), arXiv:2012.13490 [cs.LG].

[10] Momchil S Tomov, Van Q Truong, Rohan A Hundia, and Samuel J Gershman, "Dissociable neural correlates of uncertainty underlie different exploration strategies," Nature Communications **11**, 2371 (2020).

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

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

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

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

[15] 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).

[16] 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).

[17] 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).

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

[22] 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).

[23] This is true of either Monte Carlo or boot-strapped estimators (such as temporal difference learning), which then typically learn slowly or are unstable depending on the value of the learning rate.

[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] JI Gold and MN Shadlen, "The neural basis of decision making," Annu. Rev. Neurosci. **30**, 535–574 (2007), arXiv:NIHMS150003.

[26] Daeyeol Lee, Hyojung Seo, and Min Whan Jung, "Neural Basis of Reinforcement Learning and Decision Making," Annual Review of Neuroscience **35**, 287–308 (2012).

[27] Nathaniel D Daw, "Chapter 16 - Advanced Reinforcement Learning," (Academic Press, San Diego, 2014) 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] Timothy Hanks, Roozbeh Kiani, and Michael N Shadlen, "A neural mechanism of speed-accuracy tradeoff in macaque area LIP," , 1–17 (2014).

[31] 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).

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

[39] 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).

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

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

[42] 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).

[43] 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.

[44] 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).

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

[46] Sridhar Mahadevan and Bo Liu, "Basis Construction from Power Series Expansions of Value Functions," in *Advances in Neural Information Processing Systems*, Vol. 23, edited by J Lafferty, C Williams, J Shawe-Taylor, R Zemel, and A Culotta (Curran Associates, Inc., 2010)

pp. 1–9.

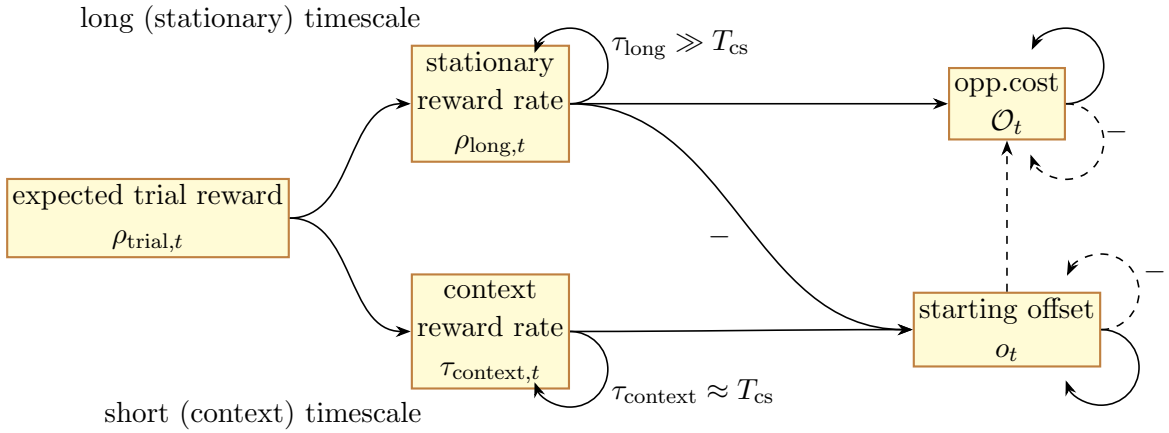

Figure 6. *Reward filtering scheme for online computation of within-trial opportunity cost.* The expected trial reward, ρ_{trial}, is integrated on both a stationary (τ_{long}) and context (τ_{context}) filtering timescale to produce estimated average and context-specific reward rate estimates, respectively. 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.

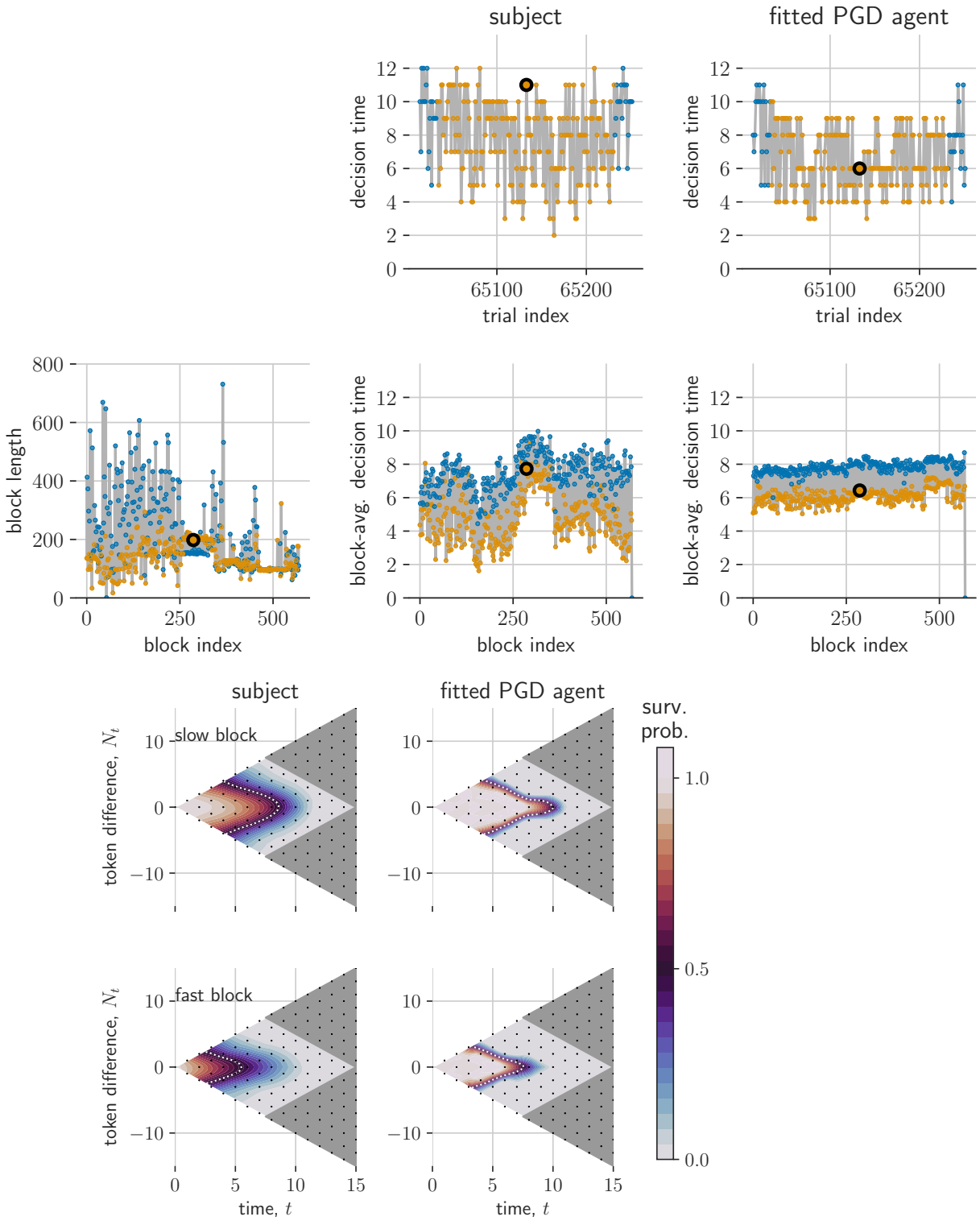

Figure 7. *Comparison of PGD and NHP in non-stationary α dynamics from [18]: Subject 2.* Same as fig. 4.

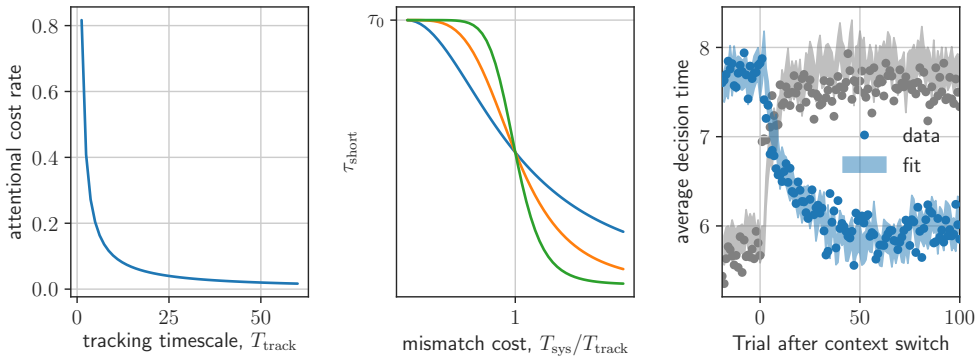

Figure 8. *Asymmetric switching cost model.* (a) Cost rate is inversely proportional to tracking timescale, T_{track}. (b) Filtering timescale τ_{context} scales down with mismatch cost $T_{\text{sys}}/T_{\text{track}}$ (sensitivity $\nu = 2, 4, 8$). (c) Adding this modified τ_{context} gives good fits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

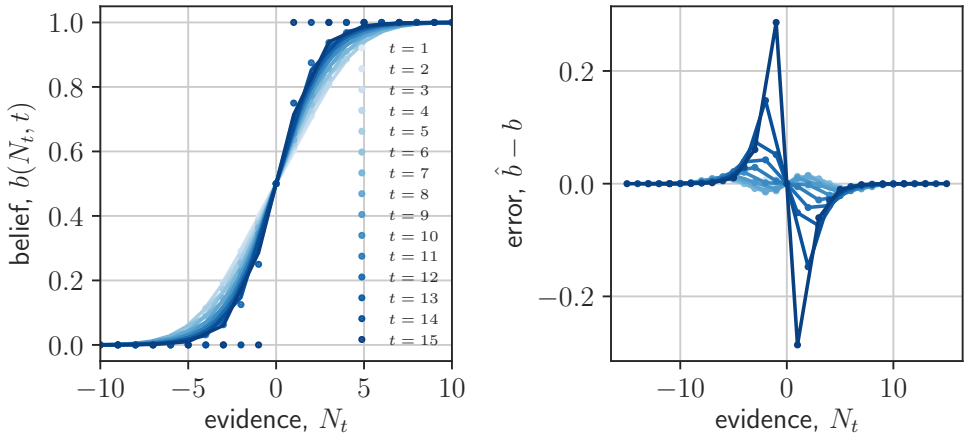

Figure 9. *sigmoidal approximation to expected reward.* (a) the approximation explained in Methods: State-conditioned expected trial reward, for different decision times. (b) The error in the approximation for different decision times.

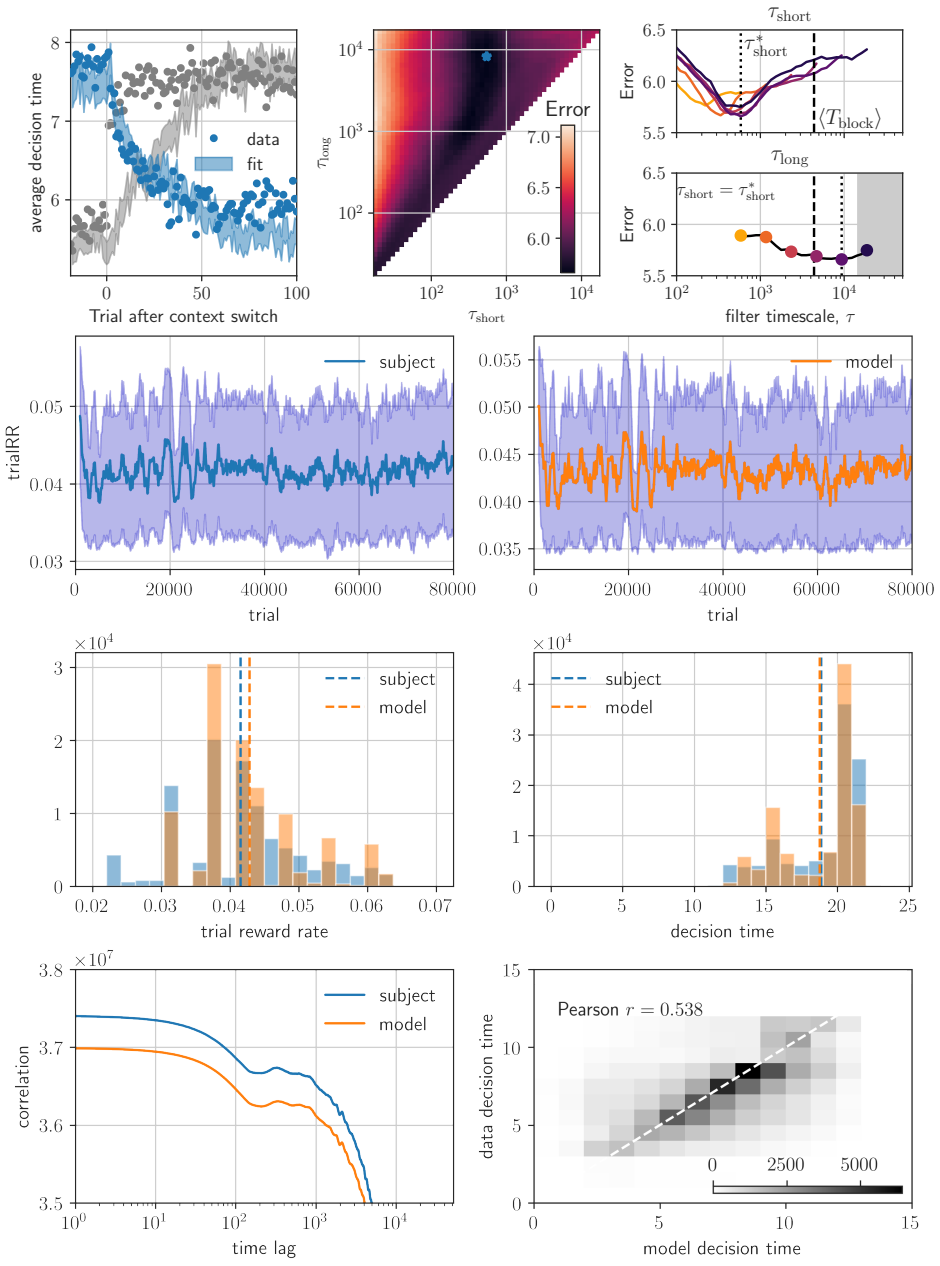

Figure 10. *Model validation on behavioural statistics from [18].* Top: Running average trial reward rate $\rho_{\text{trial},k}$ over 1000 last trials. Middle: distributions of trial reward rate (left) and decision time (right). Bottom: Auto-correlation functions (left) and cross-correlation (right: gray-scale is trial count; white dashed line is perfect correlation)

.

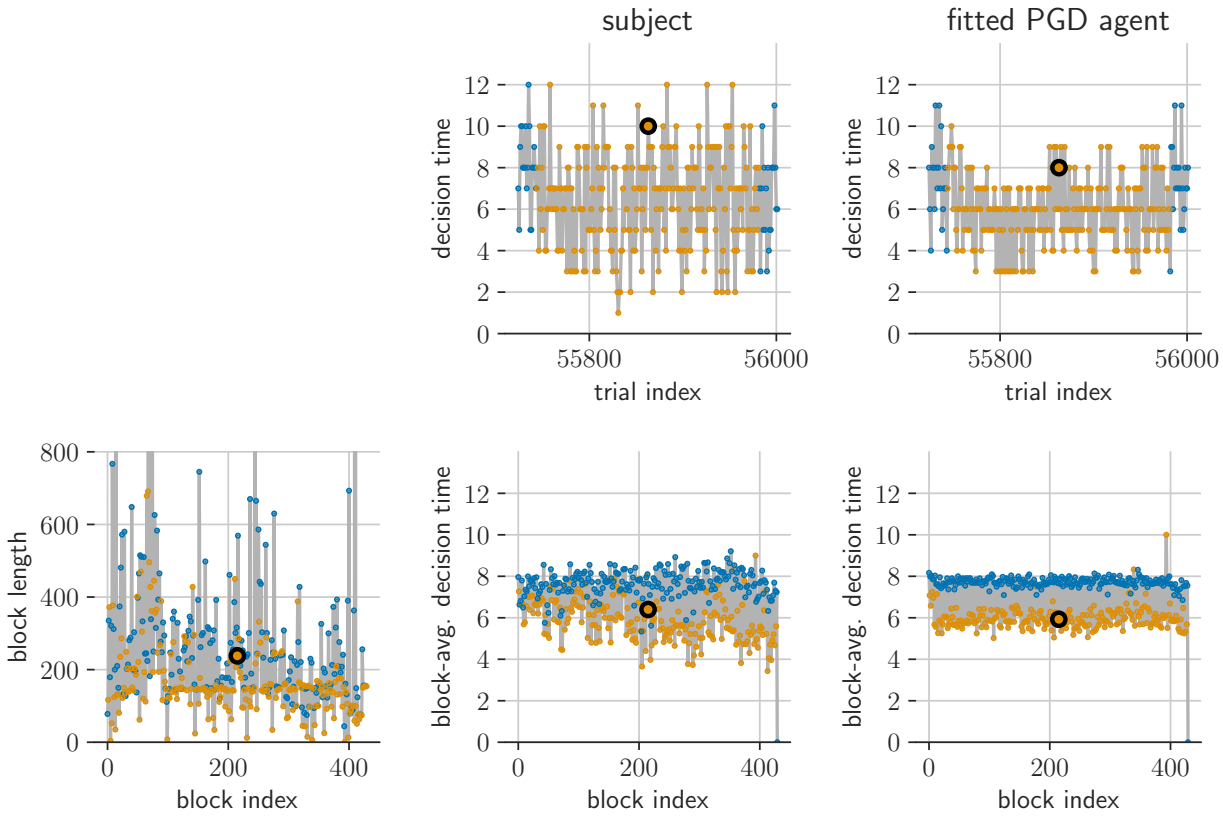

Figure 11. *Comparison of PGD and NHP in non-stationary α dynamics from [18].* (a) The sequence of trial block durations used. (b,c) Decision times during a single block. (d,e) block-averaged decision times over the experiment.

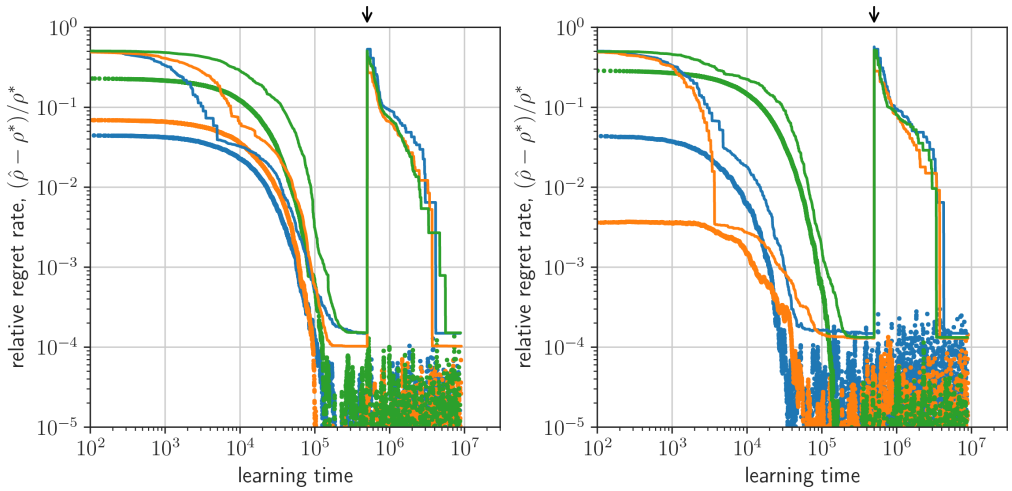

Figure 12. *Comparison of PGD and RL learning on a patch leaving task.* Performance is defined as relative regret rate, $(\hat{\rho} - \rho^*)/\rho^*$ (PGD (dots); RL (lines)). (a) Performance over different sizes of the state vector ($d = 100$ (blue), 200 (orange), 300 (green)). (b) Performance over 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.

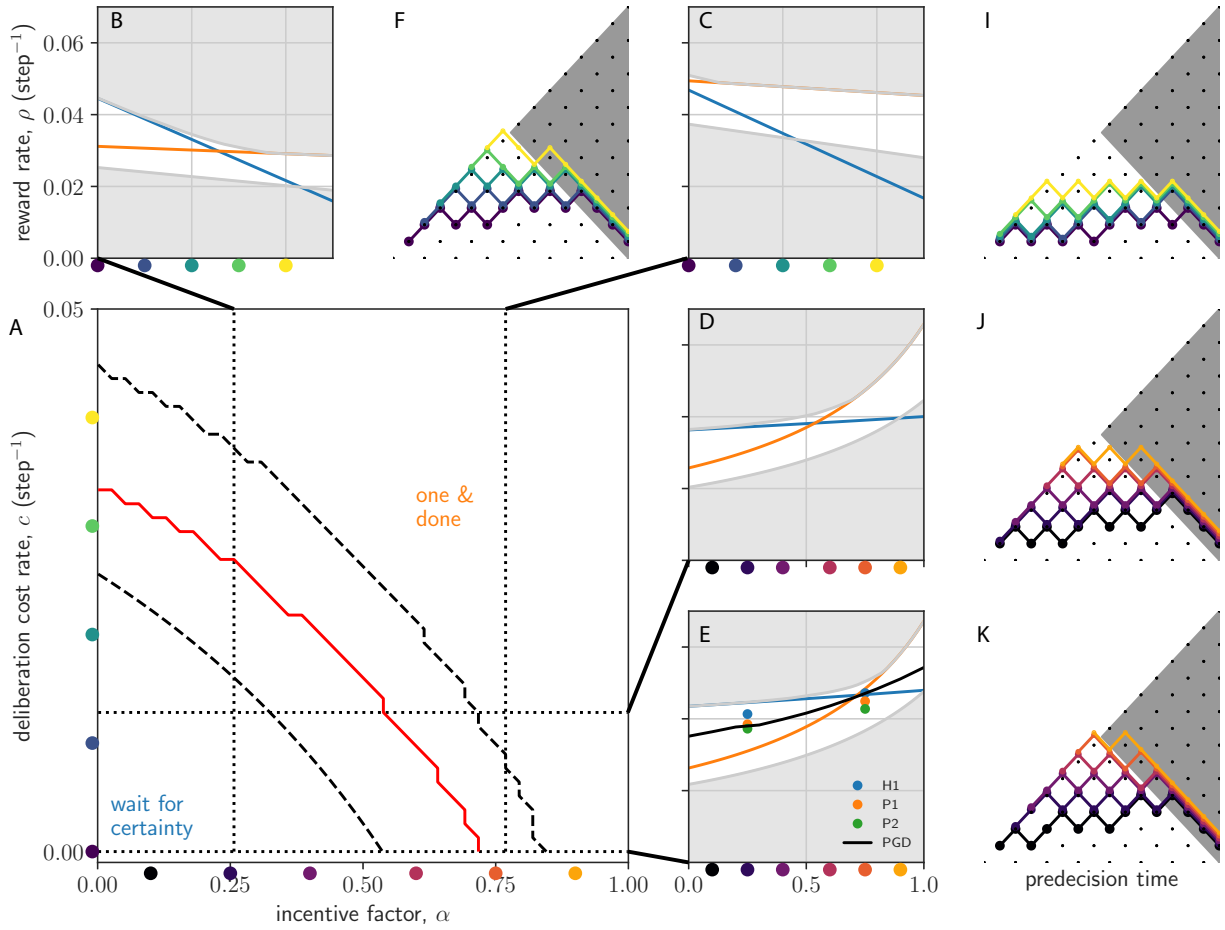

Figure 13. *Observed behaviour not in space of constant deliberation cost, reward rate-maximizing strategies.* (a) The reward-rate maximizing policy interpolates from the wait-for-certainty strategy at weak incentive (low α) and low deliberation cost (low c), to the one&done strategy at strong incentive (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&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.

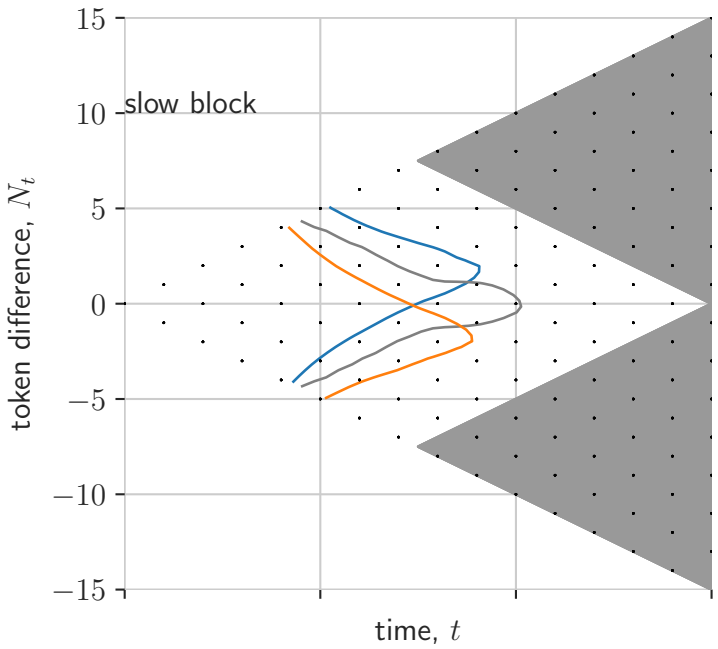

Figure 14. *Asymmetric action rewards skew survival probability.* Here, we plot the half-maximum of the PGD survival probability for three values of the action reward bias, $\gamma = -0.6, 0, 0.6$ (blue, black and orange, respectively). Other model parameters same as in fitted model.