Control Limited Perceptual Decision Making

Juan R. Castiñeiras[1†], Alfonso Renart[1†]

[1] Champalimaud Research, Champalimaud Centre for the Unknown, 1400-038 Lisbon, Portugal.

[†] Corresponding author: juan.castineiras@neuro.fchampalimaud.org
alfonso.renart@neuro.fchampalimaud.org

Bounded temporal accumulation of evidence is a canonical computation for perceptual decision making (PDM). Previously derived optimal strategies for PDM, however, ignore the fact that focusing on the task of accumulating evidence in time requires cognitive control, which is costly. Here, we derive a theoretical framework for studying how to optimally trade-off performance and control costs in PDM. We describe agents seeking to maximize reward rate in a two-alternative forced choice task, but endowed with default, stimulus-independent response policies which lead to errors and which also bias how speed and accuracy are traded off by the agent. Limitations in the agent's ability to control these default tendencies lead to optimal policies that rely on 'soft' probabilistic decision bounds with characteristic observable behavioral consequences. We show that the axis of control provides an organizing principle for how different task manipulations shape the phenomenology of PDM, including the nature and consequence of decision lapses and sequential dependencies. Our findings provide a path to the study of normative decision strategies in real biological agents.

Making the right decision often depends on specifying accurately the state of the environment. In these conditions, it is often useful to wait and gather more evidence before committing to a course of action. Indeed, organisms are able to accumulate evidence across time in order to make better decisions in the presence of sensory uncertainty[1–3]. Decades of experimental and theoretical work have shown that this process is accurately captured by the framework of bounded evidence accumulation[4,5]. While many variants of this general scheme have been developed and are used due to their ability to accurately describe choice and reaction time (RT) data from psychophysical experiments[3,4,6–10], these models are, in addition, attractive due to their normative grounding: they describe not only how agents decide, but how they ought to decide in order to satisfy reasonable decision goals.

Wald's 1945 sequential probability ratio test[11] (SPRT) provides an optimal prescription[12] for choosing online between two known alternatives, or hypothesis, during the sequential observation of the samples they produce. Using this prescription, the agent should commit when the relative evidence favoring an alternative reaches a bound, the magnitude of which measures the tolerance of the agent to making an error. The SPRT formalizes a speed-accuracy tradeoff (SAT), as accumulating enough evidence to make good decisions takes time. This is the essential computational problem in sequential sampling. Although neurally-inspired evidence accumulation models can be construed as implementing the SPRT[13], it has recently been noted that partially observable Markov decision processes (POMDPs) provide a more general and flexible normative framework for describing perceptual decisions[14–17]. MDPs are useful for defining optimal action policies in situations requiring planning, i.e., in sequential decision problems where current actions have delayed consequences[18,19]. POMDPs additionally model cases where agents are uncertain about the state of the environment and have to infer it based on noisy sensory evidence[20].

Despite these advances, a significant shortcoming of existing normative accounts of perceptual decision making is that, in their standard form, MDPs and POMDPs find optimal decision strategies that are exclusively a function of the decision problem, and are thus both context-independent and also insensitive to the particularities of the agent. In other words, the agent is viewed as a tabula rasa that can adapt its behavior with complete flexibility to the problem at hand. For real biological agents, this is a highly questionable assumption[21]. Animals face experimental tasks with existing policies and behavioral tendencies. These are very likely adaptive on evolutionary time-scales, but will be, in general, maladaptive for any particular task at hand. The context-dependent regulation of behavior requires *control*, i.e., a system for arbitrating which of several existing policies is best suited for driving action in a particular context in light of the agent's goals[22–24]. For instance, in the Stroop task[25], agents are supposed to override an existing default tendency for reporting the verbal meaning of a word and to report instead its color. Similarly, response-outcome associations or built in knowledge about the time-scales of variation of the environment typically generate sequential response dependencies in psychophysical tasks[26,27] that are maladaptive when behavior should be under the exclusive control of the stimulus.

A way forward is to acknowledge that controlling default action policies has a cost[28–30] and to include this cost in the optimization process that selects a policy[31–33]. In this view, a policy that might appear clearly suboptimal for a particular task (when only performance costs are considered), can become optimal when both performance and control costs are evaluated. An attractive framework for exploring this problem is Kullback-Leibler (KL) control[34–37]. In KL control, agents

are assumed to possess a stochastic default policy and the immediate cost of an action under any candidate policy contains a term that grows with the dissimilarity between the likelihood of the action in that state under the default and candidate policies. This implements a cost of control. The specific form of the control cost both facilitates computation[35,38] and leads to optimal policies with desirable properties in terms of information seeking and complexity[36,39] and efficient use of limited computational resources[40].

Here, we characterize the consequences that derive from a trade-off between performance and control costs on perceptual decision-making. In particular, we extend the KL control framework to describe optimal control of point actions in continuous time under state uncertainty. Since the essential requirement for making good perceptual choices is the ability to wait until enough information has been gathered, we consider the effect of default policies that embody a certain probabilistic tendency to respond per unit time independently of the stimulus during evidence accumulation, thus biasing the SAT and, for biased default policies, also the choice preferences of the subject. We show that control limitations are associated to optimal decision policies that rely on 'soft' decision bounds, and describe how these limitations shape reward rate, choice accuracy, RT and decision confidence (DC). We identify behavioral signatures of decision-making that are characteristic of the control-limited regime, and show that this regime is expected to be found in conditions of high time-pressure and for easy discriminations, in agreement with experimental findings. In this way, variations in control provide an organizing principle for a wide range of observed but previously unaccounted for behavioral observations from a normative perspective. Finally, we show how to correctly recover the true psychophysical ability of subjects in the presence of lapses caused by limitations of control, and how to identify different targets of across-trial choice dependencies.

Control limitations shape action values and decision policies

We consider a classic decision-making paradigm in which an agent makes a binary choice about a latent variable based on a stream of stochastic sensory observations. The solution to this task involves a decision about when to commit to one of the two options. However, before studying this problem, we schematically explain the ingredients of the framework for control-limited decision making using a simple example involving a decision at a single time-step without sensory uncertainty. In this example (Fig. 1a), there are varying numbers of visual stimuli on either side of the long arm of a T-maze, and a rat would need to make a choice at the decision point by turning towards the side with more stimuli in order to obtain a larger reward. We assume the rat has default tendencies to perform these same turning actions – such as a propensity towards spatial alternation (Fig. 1a). Following a spatial alternation policy will lead to less reward, but overriding this default tendency requires control and control is costly[25,28–30]. Consider the two consecutive trials in Fig. 1a. Based on the sensory input and on the contingencies of the task, Left is the correct response in trial $n+1$, so this response would have a higher (raw) action value $R(a,s)$, where a is an action and s is an environmental state (Fig. 1b*i*). At the same time, the rat's tendency to alternate furnishes a default action policy $P_d(a|s)$ which, given that Left was chosen in trial n, assigns higher probability to the Right (alternating) response in the current trial (Fig. 1b*ii*). How should these two tendencies be traded off? In the KL control framework[34–36], one seeks a policy $P_o(a|s)$ that minimizes a total cost (hence, an 'optimal' policy) equal to the weighted sum of a performance cost and a control cost

$$\text{Total Cost}\,[P_o; P_d] = \text{Performance Cost}\,[P_o] + \frac{1}{\beta}\,\text{Control Cost}\,[P_o; P_d] \quad . \tag{1}$$

The control cost takes the functional form

$$\text{Control Cost}\,[P_o; P_d] = KL\,[P_o|P_d] \tag{2}$$

where $KL\,[P_o|P_d]$ is the Kullback-Leibler divergence between the optimal and default policies – a quantitative measure of the dissimilarity between two probability distributions[41] (Supplementary Information). The performance cost is the standard objective function in a sequential decision problem (for instance, the negative future discounted reward following a given policy).

The control aspect of the problem has thus two elements. First, the term $KL\,[P_o|P_d]$ measures the amount of conflict that the task induces for the agent. If the default response tendencies are aligned with the current task demands, it will be close to zero, whereas it will be large if the two are inconsistent, as in the example above where reward-maximizing and default policies can favor different actions in the same state (Fig. 1a, bi-ii). Second, the relevance of this kind of conflict in shaping the behavior of the agent is given by the positive constant β, which determines the relative importance of performance and control costs. When β is very large, the total cost is effectively just determined by performance considerations regardless of how much conflict the task induces, which can be understood as the agent being able to muster the required control to adapt its behavior. At the other extreme, values of β close to zero represent agents for which modifying default behavior is extremely costly and which will thus use essentially the same default response strategy regardless of the task at hand. For intermediate values of β, agents display varying degrees of adaptability and control. In light of this, we will refer to β as the 'control ability' of the agent and to β^{-1} as its 'control limitation'. When $\beta^{-1} = 0$, the agent displays fully adaptable behavior, so we will refer to this as the FAB agent.

In a standard setting ignoring control, the optimal policy would select the action in a particular state that maximizes value[18,19] (minimizes cost). In the simple one-shot problem in Fig. 1, the rat would pick the action for which $R(a, s)$ is larger (in a sequential decision problem, the value of the state would include long-term consequences). It can be shown (Supplementary Information) that KL control modifies this picture in two ways. First, control-limited agents use policies that are probabilistic[35,36]. Instead of maximizing, i.e., of choosing the action with the largest action-value, actions are chosen using a soft-max rule parametrized by β (Fig. 1b*v*, Supplementary Information). Control limitations are thus associated to exploration. Exploration in the KL framework is self-consistent, in the sense that action-values are computed using the optimal stochastic policy, and the soft-max rule is not *ad hoc*, but a necessary consequence of measuring control costs using the KL divergence[42]. The second modification introduced by the KL framework is that exploratory policies are biased towards the default. It can be shown (Supplementary Information) that the raw action values $R(a, s)$ need to be redefined to new quantities $\tilde{R}(a, s)$ given by

$$\tilde{R}(a, s) = R(a, s) + \beta^{-1} \log[P_d(a|s)] \tag{3}$$

Thus, the raw action-value of choosing action a in state s is reduced according to how surprising it is that the agent would choose that action in state s following the default policy, with the total reduction being proportional to the control limitation of the agent β^{-1}. In our example in Fig. 1, the

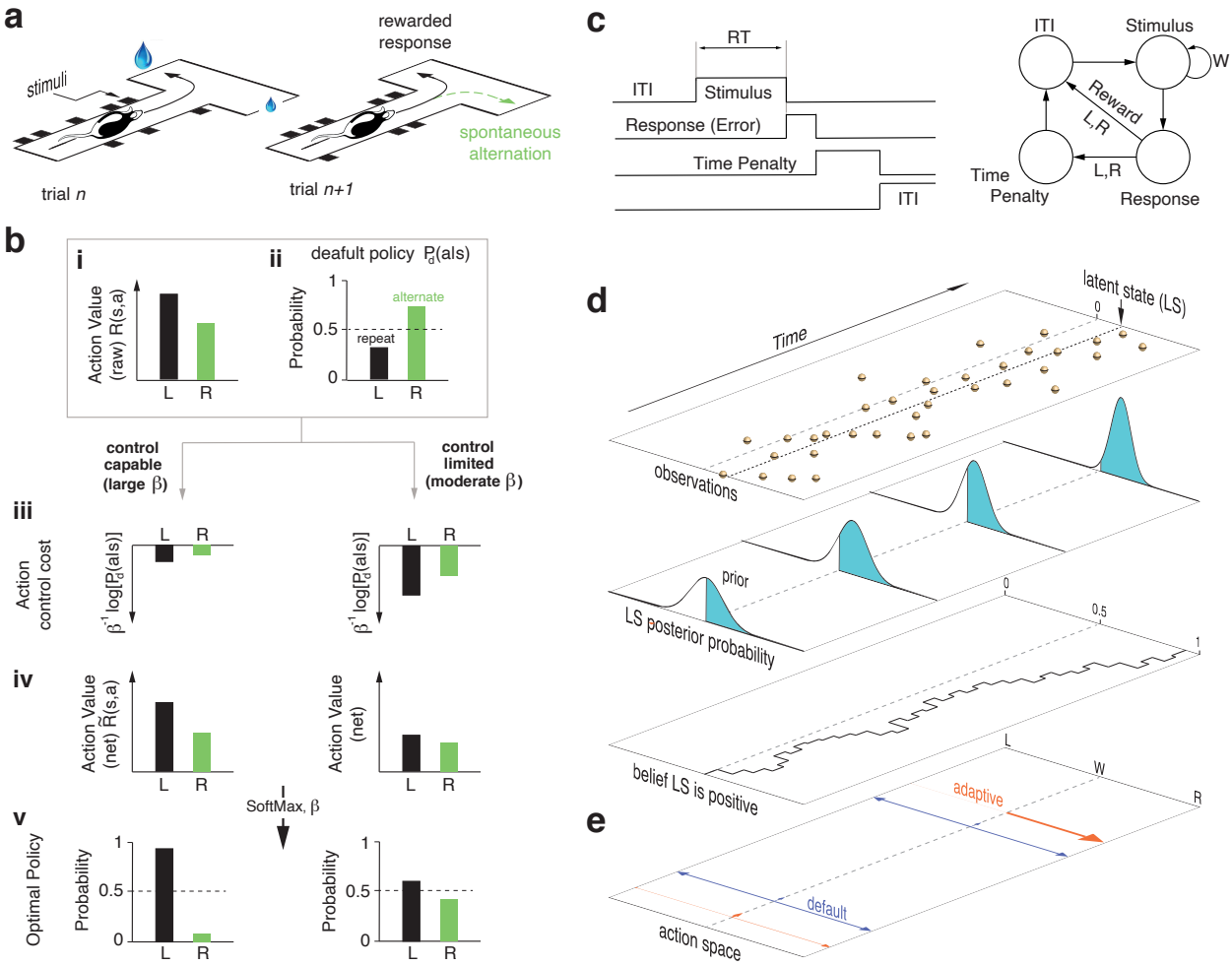


Fig 1. Formalism for control limited perceptual decision making. **(a)** Example of a perceptual decision problem where control limitations are relevant. A rat needs to make a turn at the decision point towards the side where more visual stimuli were displayed in the long arm of a T-maze. The rat has a default tendency to alternate, which may conflict with the task-appropriate response in some trials, like the trial $n+1$ on the right. **(b)** Ingredients of the framework shaping behavior at trial $n+1$. Raw action value $R(s,a)$ (i) and default policy for this trial (ii) L, R stand for left and right respectively. $(iii-v)$ Consequences of control limitation for a control capable (left) and a moderately control limited agent (right). Control action cost (iii) and effective (net) action value $\tilde{R}(s,a)$ (iv), equal to the sum between i and iii. (v) The control capable agent effectively maximizes (chooses the action with higher value), whereas the control limited agent chooses according to a weighted average. **(c)** Trial-structure of the tasks we study. Left. The time during which the stimulus is sampled before a choice is made defines the reaction time (RT) of a trial. Depending on the outcome, a point reward can be earned (for correct choices) or a time penalty is imposed (for errors). There is an inter-trial interval (ITI) before the next stimulus is presented. Right. State space with possible transitions in the task. W stands for wait. **(d)** Schematic description of the inferential process in one trial of a sequential sampling problem. The agent experiences noisy observations whose mean (black dotted line) depends on the value of a continuous latent state (top). The task of the agent is to decide if the latent state is positive or negative. To do this, it first updates its prior belief about the value of the latent state based on the incoming information, into a posterior distribution. As more information is sampled throughout the trial, the mean of the posterior approaches the true mean and the posterior uncertainty decreases (middle). In a categorical binary decision problem, outcomes depend only on the belief that the latent state is positive (the area under the posterior on the positive side; bottom), which fluctuates stochastically driven by the noisy observations. Observations come at discrete time-steps in this example for illustration purposes, but time is continuous in our model.

(e) At each time, the agent chooses between committing to either of the two possible options (left/right), or waiting and accumulating more evidence. Unless specified otherwise, we typically consider unbiased default policies with a constant probability per unit time of making a random left/right choice. Adaptive strategies update the probability of each action using the agent's task-relevant belief.

net value of both actions becomes more similar for a moderately control-limited rat, because the raw action value and the cost of control lead to different preferences (Fig. 1b*iii-iv*). KL control provides thus a particular instantiation of optimal policies that include both directed (towards the default) and random (as quantified by the agent's control limitation β^{-1}) components of exploration[43]. Since both of these components are parametrized by the control ability β (but see Supplementary Information for a relaxation of this assumption), optimal control-limited policies converge to the standard (deterministic) optimal policy for the same task when β^{-1} approaches zero.

We model the structure of a typical decision-making experiment in the laboratory, with one binary decision per trial, difficulty varying randomly across trials, a time penalty for errors, and an inter-trial interval (Fig. 1c). Because trials are typically short, but there are many of them in one session, we assume that the goal of the agent is to maximize the 'reward rate' across the whole session (see Supplementary Information for details on how to use this performance measure in MDPs). This allows the description of agents sensitive to the long-term consequences of their actions without the need to invoke temporal discounting[44,45]. We emphasize again that control costs are included in this optimization, so that what the agent is actually maximizing is the negative total cost in Eq. (1) per unit time, which we denote as ρ.

By definition, the relevant state of the environment in a perceptual decision-making problem is latent (not directly observable), but it can be inferred using stochastic observations. Without loss of generality we assume that the latent state is continuous and equal to μ, and emits temporally uncorrelated Gaussian observations with mean μ and variance $\sigma^2 dt$. For a categorical binary choice, the task can be cast as that of deciding about the sign of the latent state[15]. The absolute magnitude of the latent state defines the strength of the evidence for a given decision, which we assume is drawn randomly from trial to trial from a Gaussian prior distribution with mean zero and variance σ_μ^2. The agent begins the trial undecided and with the correct prior over the value of the latent state, and uses the observations to sequentially update the posterior probability over μ. Since the task only requires a report on the sign of the latent state, all future consequences depend only on the agent's belief that the latent state is positive (Fig. 1d), which we denote by $g(t)$. This recursive inferential process defines a stochastic trajectory on $g(t)$ (Fig. 1d, bottom, Fig. 2c), which can be mapped one-to-one from the stochastic trajectory of accumulated evidence in each trial[15] (Supplementary Fig. 1).

At each point in time, the agent can choose between committing to either of the two possible options (which we describe as Left (L) and Right (R), with R being correct when the latent variable is positive), and continuing to sample the stimulus (i.e., 'waiting' (W), Fig. 1e). We consider agents with default response policies that are not task-adapted, and which will thus need to exercise control in order to perform well. By default, agents have a certain constant probability of commitment λ per unit time and, when they do commit, select one of the two options randomly (with or without bias). These default policies describe agents with a propensity to lapse in all trials, with an exponential RT distribution of mean λ^{-1} s. The problem we describe can thus be cast as an agent trying to maximize reward given a limited ability to control a default tendency to lapse with a certain

urgency throughout the trial. Adaptive behavior in the task requires the agent to control these default policies in two ways. First, the tendency to respond needs to be matched to the requirements of the task. For instance, if the task emphasizes accuracy over speed but λ is large, control will have to be used to slow responding down. The magnitude of the parameter λ can thus be understood as biasing the agent towards or against speed in the speed-accuracy tradeoff. Second, the actual categorical choice needs to become stimulus-dependent, i.e., response probability needs to become a function of $g(t)$ (Fig. 1e).

Optimal policies for control-limited agents consist of smooth decision bounds

Previous studies showed that, in the absence of control limitations, the optimal policy is for the agent to make a choice when its instantaneous belief $g(t)$ reaches a certain time-dependent decaying bound[14,15]. This bound corresponds to the moment where the action-value of committing to either of the two options – which tends to grow through the trial – becomes equal to the initially larger long-term value of the uncommitted state – which decays (Supplementary Fig 1). Since this bound on belief corresponds to a bound on accumulated evidence (albeit with a different shape; Supplementary Fig. 1), the optimal policy has a neural implementation in terms of a drift-diffusion model[14,15,46], as long as the temporal evolution of the decision bounds can be specified accurately.

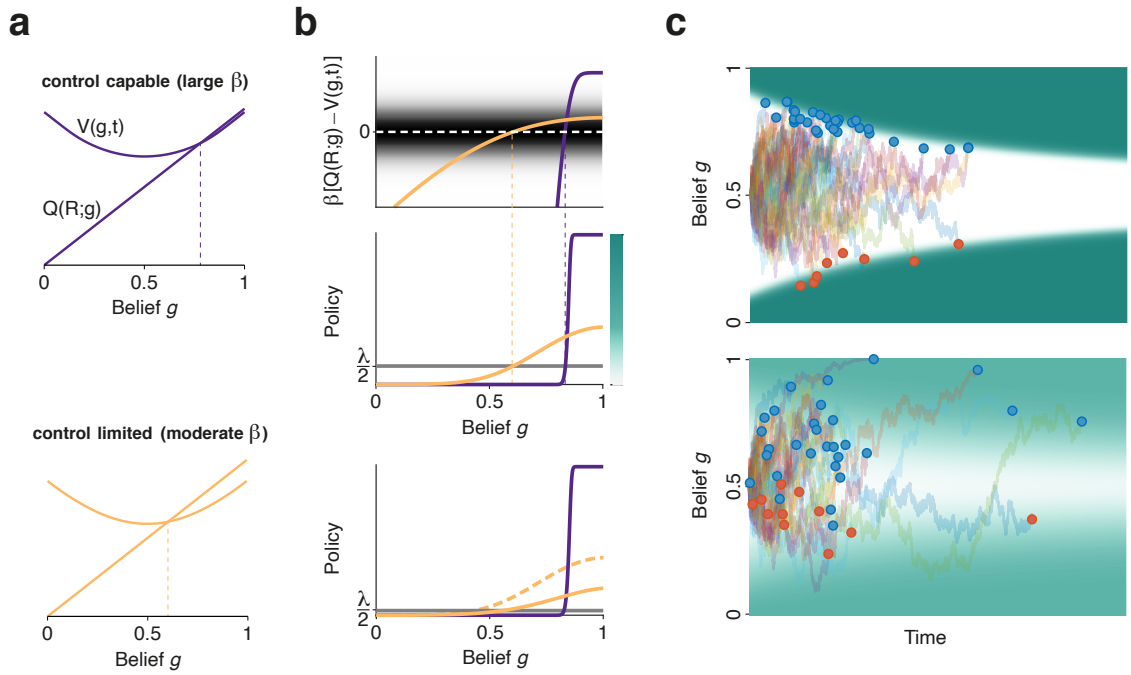


Fig 2. Effect of control limitations on optimal decision policies. (a) Value function for waiting (convex curve) and action value for choosing R (straight line) as a function of belief at the beginning of the trial ($t = 0$), for a control capable (top) and a control-limited agent (bottom). Dashed lines indicate the belief for which both quantities are equal. **(b)** Top. Difference between the two curves in (a) for each agent (same color code as in (a)) scaled by the agent's control ability β. Dark background signals the region where the scaled difference in action values is of order unity. The probability of committing to a choice varies over the range of

beliefs for which the scaled difference overlaps with this region. Middle. Choice policies at this time for the two agents. Vertical dashed lines mark the transition from suppressing to promoting choice relative to the default policy $P_d(R) = \lambda/2$. Bottom. Same as middle but for agents where the default response rate has been reduced by a factor of 4. For comparison, the policy for the control-limited agent in the panel above is also shown here as a dashed line. **(c)** Decision-making dynamics for control capable (top) and limited (bottom) agents, with the same default response rate λ as in (b), top. Traces are belief trajectories (Methods) for a set of trials with moderate strength of evidence towards the right (i.e., positive beliefs). Background represents the policy (instantaneous probabilities of choice commitment per unit time; see colorbar in (b), bottom, for the relative magnitudes of these probabilities). Dots signal the moment of commitment (blue, correct; red, error trials). Parameter values for all figures are listed in Supplementary Table 1.

In order to investigate optimal decision policies in control-limited agents, we extended the KL formalism to incorporate sensory uncertainty (partial observability) and continuous states (for the agent's belief) and discrete actions in continuous time (Supplementary Information). Our results show that optimal control-limited policies generalize naturally the policy just described: instead of transitioning discontinuously from waiting to responding, they are described by a temporally evolving probability of commitment which, for any fixed time, grows with the action-value of the two options relative to the value of the uncommitted state (Fig. 2), with the steepness of the transition growing with β. In fact, the optimal policy is given by a simple mathematical expression (Supplementary Information)

$$P_o(R, L|g, t) = P_d(R, L)\ \exp\Big(\beta[Q(R, L; g, t) - V(g, t)]\Big) \tag{4}$$

Here, $Q(a; s)$ is the action-value of action $a = R, L, W$ in state $s = g, t$, and $V(s)$ is the long-term value of the uncommitted state. The exponential term provides a gain factor on the agent's default probability of committing to either action per unit time $P_d(R, L) = \lambda/2$, and the belief at which the action-value of commitment and the value of the uncommitted state become equal marks a transition from suppressing the default tendency to respond to augmenting it (Fig. 2b, top). In control-capable agents, the transition from low to high probability of commitment as a function of belief $g(t)$ is sharp, and the probability per unit time in the high state is large (Fig. 2b, middle), effectively resembling a hard bound on $g(t)$ (Fig. 2c, top). In contrast, the more control-limited an agent is, the more similar the commitment probabilities in the low and high states become, and the larger the range of beliefs over which the transition occurs (Fig. 2b, middle). For such agents, behavior is more stochastic and there are broad ranges of belief and time for which the agent can either be committed or uncommitted in different trials (Fig. 2c, bottom). When the default rate of responding λ of the agent changes – for instance, decreasing – the probability per unit time of commitment as a function of belief is scaled down, but only for control-limited agents (Fig. 2b, bottom). This is equivalent to a default emphasis on accuracy (over speed). In sum, optimal decision policies for control-limited agents with lapsing default policies resemble smooth decision bounds. This is a form of noise-induced linearization, a common phenomenon which takes place in many physical systems, including neurons[47].

The task stakes specify the behavior of the FAB agent

Before examining the effect of control limitations on the phenomenology of decision making, we consider a more general question, namely, how large is the space of possible optimal solutions to a

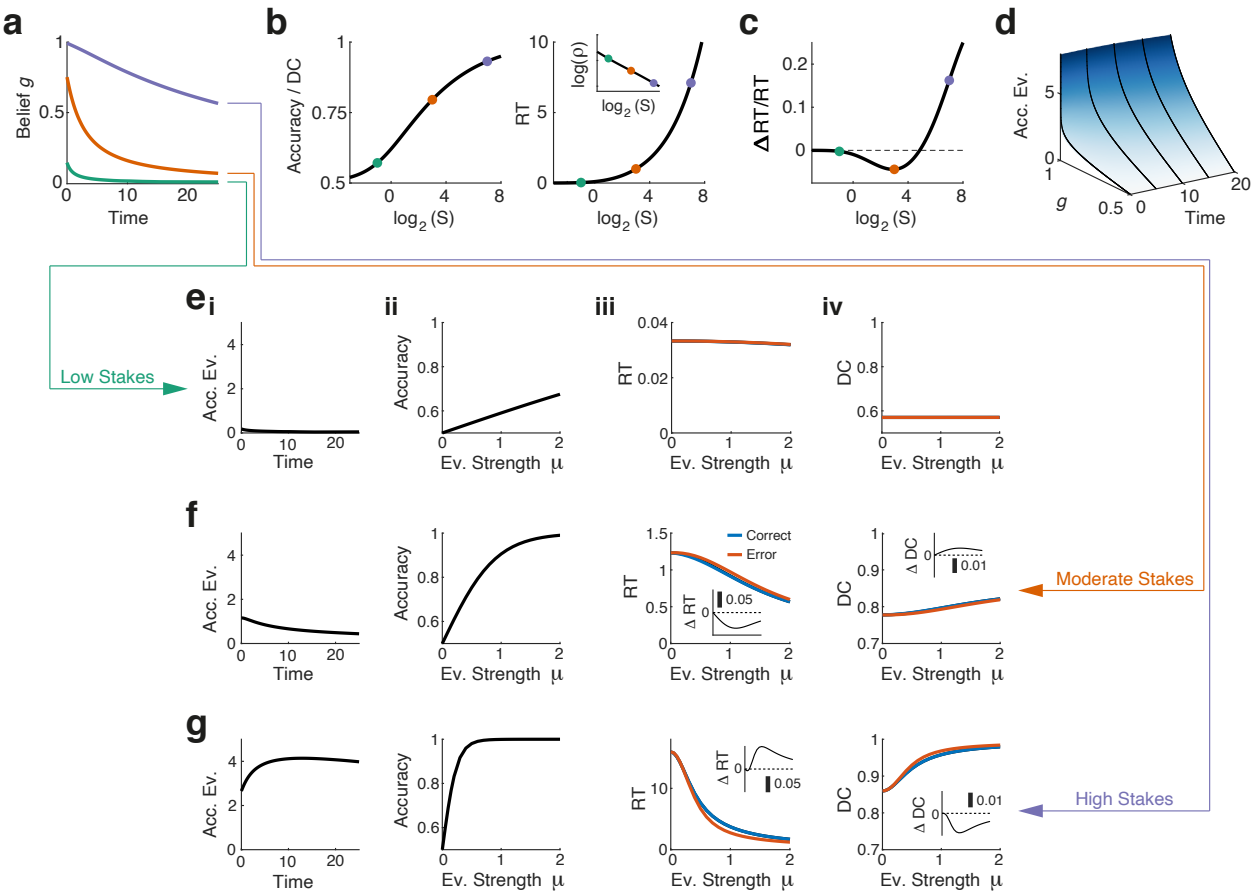


Fig. 3. Decision policies are determined by the task stakes. All panels in this figure describe the behavior of the FAB agent (i.e., $\beta^{-1} = 0$). **(a)** Decision bounds on belief for three values of the stakes S (see text). **(b)** Left. Accuracy averaged across all difficulties (equal to the decision confidence (DC)[48]) as a function of the task stakes S. Right. Same for RT. Inset. Reward rate decreases monotonically with S. The three values of S in (a) are marked with circles of corresponding colors in (b,c). **(c)** Relative difference in (average) RT between correct and error trials as a function S. **(d)** One to one map between instantaneous belief g and time and accumulated evidence during the trial. **(e)** Behavior of the agent for a low-stakes task. *i*. Decision bound on accumulated evidence. *ii*. Psychometric function. *iii*. Chronometric functions for correct (blue) and error (red) trials. *iv*. DC as a function of evidence strength for both outcomes. In panels ($ii - iv$) the upper limit on the strength of evidence μ is twice the width σ_μ of the prior. **(f,g)** Same as (e) but for moderate and high stakes respectively. Insets in (*iii*, *iv*) show difference in RT and DC respectively between correct and error trials.

sequential sampling decision problem? Although the problem as we have construed it depends on more than a handful of parameters (specifically seven: five for the task – the noise in the stimulus σ^2, the width of the prior σ_μ^2, the reward magnitude R_w and penalty time t_P, and the inter-trial-interval t_{ITI} – plus two for the agent – the control ability β and the default response rate λ) it can be shown that the task faced by the agent depends effectively on a single dimensionless parameter This parameter, which we refer to as the 'stakes', is given by $S = (t_P + t_{\text{ITI}})/t_g$, i.e., the sum of the penalty time and the inter-trial interval relative to $t_g = \sigma^2/\sigma_\mu^2$, which describes the intrinsic time-scale of the inference process. t_g measures the stimulus sampling-time that it takes the agent to reduce by one half its initial uncertainty about the strength of evidence. Unless noted, we always

measure constants with units of time or rate with respect to t_g (i.e., $t_g \equiv 1$). Intuitively, when $S \gg 1$, stimuli are presented rarely, so maximizing the reward rate demands that the agent samples the stimulus sufficiently long to make accurate choices, i.e., the stakes for the agent for performing well in the task are high. Conversely, when $S \ll 1$, stimuli arrive so frequently that it becomes worthless for the agent to invest time in sampling the stimulus, as another opportunity will be presented soon in which reward can be obtained with at least 50% probability. Because the magnitude of the stakes determines in this fashion the optimal stimulus sampling time allocation, one can think of the stakes as quantitatively specifying the speed-accuracy demands associated to the task.

For the FAB agent ($\beta^{-1} = 0$), the optimal policy depends exclusively on the stakes[4,49] (Fig. 3). It is instructive to understand this situation in detail, as sweeping the value of the stakes defines the whole universe of optimal solutions to the sequential sampling decision problem. The agent adapts its policy to the stakes of the task by raising the decision bound when accurate performance is needed (Fig. 3a). As this happens, both accuracy and RT naturally grow (Fig. 3b). Reaction time grows faster which, together with the larger interval between trials, results in a monotonic decrease of the reward rate of the policy with the stakes of the task (Fig. 3b, inset). The fact that agents solve the task by deciding how much time to allocate to each stimulus according to the background rate of stimulus presentation in the environment suggests interesting connections between perceptual decision making and foraging theory[50].

A number of quantitative signatures of behavior have been identified as useful for distinguishing between different mechanistic implementations of the decision making process. One of them is the difference in RT between correct and error trials[51]. This difference is non-monotonic in the task stakes (Fig. 3c), defining three qualitatively different regimes in this problem. For low enough stakes, it is negligible. For intermediate stakes, incorrect decisions take slightly longer than correct ones, and for sufficiently high stakes this pattern is reversed. The reversal is a consequence of the shape of the map between accumulated evidence and belief $g(t)$ (Fig. 3d, Supplementary Fig. 1). It is known that decaying bounds result in larger RTs for errors[52]. However, it is important to realize that the relevant bounds are those on accumulated evidence, not on belief. Although the bounds on belief always decay with time (Fig. 3a), extremely large values of accumulated evidence are necessary to reach large values of $g(t)$, specially at long times (Fig. 3d), resulting in a situation where the optimal bounds on accumulated evidence initially grow with time when the stakes are sufficiently high (Fig. 3g, Supplementary Fig. 1). Empirically, the relationship between the RTs of correct and error trials is seen to vary with the specific conditions of the discrimination task[8,53,54], as we discuss below (Fig. 6).

Decision making in control-limited agents

For control-limited agents, the task depends effectively both on the stakes and the inter-trial interval , although the phenomenology does not change qualitatively from this extra parameter. We generally fix $t_{\mathrm{ITI}} = t_g = 1$ and control the stakes S by manipulating the time penalty for errors t_p (i.e., $S = t_p + 1$). In addition, the optimal policy now depends also on the properties of the agent, both its default response rate λ as well as its capability for control β. In Fig. 4 we show how the main features of the behavior of the agent in the discrimination task depend on these three parameters. When the control ability β is sufficiently low, the agent behaves essentially according to the default

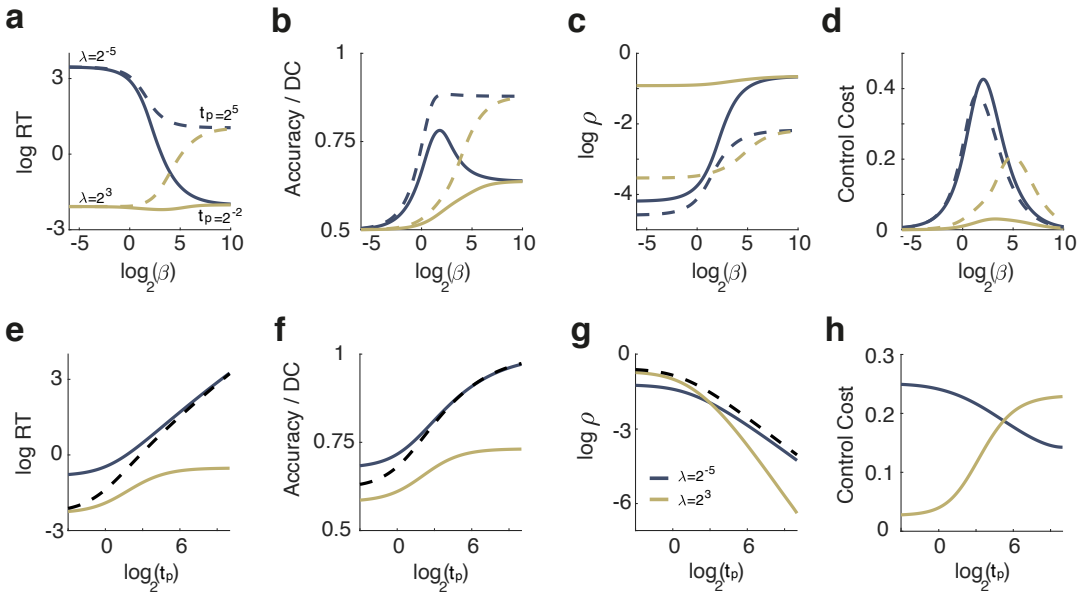


Fig. 4. Perceptual decisions in control-limited agents. (a) RT (averaged across difficulties) as a function of the control ability of the agent β. Blue and orange represent agents with low and high default response rates, respectively. Dashed and solid lines represent tasks with high or low t_p, respectively, which result in correspondingly high or low stakes. **(b-d)** Same as (a) but for accuracy/decision confidence (b), reward rate ρ (c) and the total control cost (d) respectively. **(e-f)** Same four quantities, but plotted as a function of the penalty time t_p (i.e., stakes) for a moderately control-limited agent ($\beta = 2^4$). For comparison, in (e-g) the black dashed line shows the behavior of the FAB agent. Because we are fixing $t_g = 1$, throughout this figure $S = t_p + 1$.

policy. The mean RT is thus equal to λ^{-1} and accuracy is at chance level independently of the task stakes (Fig. 4a-b). At the other extreme, when β is sufficiently high, the agent's behavior is unaffected by the default policy (Fig. 4a-b) and depends only on the stakes (Fig. 3). Accuracy is non-monotonic with β if the stakes are low and the agent is accuracy-biased ($\lambda < 1$). This happens because accuracy initially increases as the agent becomes able to adapt its behavior to the task. Because this agent underemphasizes speed, its accuracy can grow to be quite high even for a moderate increase in β. However, because the stakes are low, the fully adaptive strategy is to forego of accuracy and decide quickly instead (Fig. 3b), leading to the non-monotonic behavior in Fig. 4b.

The reward rate ρ always grows with the control ability β (Fig. 4c), confirming the intuition that control-limitations always represent a handicap for the agent. The monotonic relationship between β and the reward rate establishes an exploitable link between control and motivation[32,55], which we address below (Fig. 6; Discussion). The total control cost, given by $\beta^{-1}KL[P_o|P_d]$, is always zero for extreme values of the control ability. When β is close to zero, control is too costly for the agent, so the optimal strategy is to operate under the default policy, in which case the KL term vanishes. At the other extreme, when β^{-1} is near zero, the total control cost is zero because there are no control limitations. The control cost is maximal at intermediate values of β (Fig. 4d).

One can also look at the behavior of the agent as a function of the stakes when the control ability is fixed. This reveals explicitly that, compared to the FAB agent, the control-limited agent is not able to adapt its behavior to the demands of the task. For instance, an accuracy-biased agent can have similar accuracy and RT as the FAB agent, but only when the stakes are very high (Fig. 4e-f), since in those conditions its default RT is well-matched to the demands of the task. The delayed

commitment of this agent allows it to actually outperform the FAB agent in terms of accuracy when the stakes are low (Fig. 4f), but the extra time invested does not pay off sufficiently, leading to a suboptimal reward rate in (Fig. 4g). When the stakes tend to require RTs matched to the default commitment rate of an agent, the corresponding control costs under the optimal policy are smaller (Fig. 4h).

Signatures of control limitations on decision confidence

The smoothing of the decision bounds caused by the control-limitations of the agent (Fig. 2) strongly shapes the transformation from sensory evidence into categorical choices. In particular, limitations in control alter the beliefs of the agent at the moment of commitment, i.e., the agent's decision confidence (DC). DC in a categorical choice measures the decision-maker's belief in her choice being correct[56,57]. A research program within psychology and cognitive neuroscience has documented how explicit judgements (or implicit measures[58]) of DC depend on various properties of the decision problem, such as discrimination difficulty, trial outcome or time pressure[56,59–61]. Normative approaches have also explored the phenomenology of DC that follows from statistically optimal decision strategies[62–64].

In order to systematically characterize how control-limitations shape DC, we examined the behavior of a control capable ($\beta = 2^8$) and a moderately control-limited ($\beta = 2^{3.5}$) agent for easy and difficult discriminations. In both discriminations the latent variable is positive (i.e., R is the correct choice), but we varied the strength of the evidence. We recall that $g(t)$ describes the agent's belief that the latent variable is positive. Thus, DC is given by the value of $g(t)$ at the moment of commitment for rightward decisions, and by $1 - g(t)$ for leftward ones. When the strength of evidence is weak, individual trials are compatible with beliefs spanning both choice options depending on the stochastic evidence (Fig. 5a left), whereas when the strength of evidence is large, the agent's beliefs quickly converge on rightward preferences (Fig. 5a right). We developed methods to compute semi-analytically the joint distribution of DC and RT corresponding to the optimal policy of any of our agents (Supplementary Information).

For control-capable agents, this distribution is tightly focused (Fig. 5b) around the regions where the probability of commitment abruptly transitions from zero to a large value (Fig. 2b,c), which approximates the policy of the FAB agent, based on a temporally decaying decision bound[15] (Fig. 3). When the strength of evidence grows, the distribution still tracks the symmetric bounds but is shifted towards earlier RTs, and thus more extreme beliefs (Fig. 5b, right; Fig. 5c, left). Because the bounds are necessarily outcome-symmetric, DC is almost outcome-independent and grows with the strength of evidence (Fig. 5c, right). The outcome-dependence of DC is referred to as confidence resolution (CR)[56,59,60]. Interestingly, the optimal decision policy without control limitations, which is always superior in terms of reward rate, has poor CR[17]. A tendency for decision confidence to increase with the strength of evidence regardless of outcome is sometimes observed[62,65,66], specially in experiments requiring simultaneous reports of decision confidence and choice[62,65] (Discussion).

For the control-limited agent, the joint distribution of DC and RT is much less concentrated (Fig. 5d; see also Fig. 2c). When the strength of evidence is almost zero, the distributions for correct and error trials are still approximately symmetric (Fig. 5d, left). But for easy conditions both distributions shift up towards rightward beliefs (i.e., towards the evidence), making them

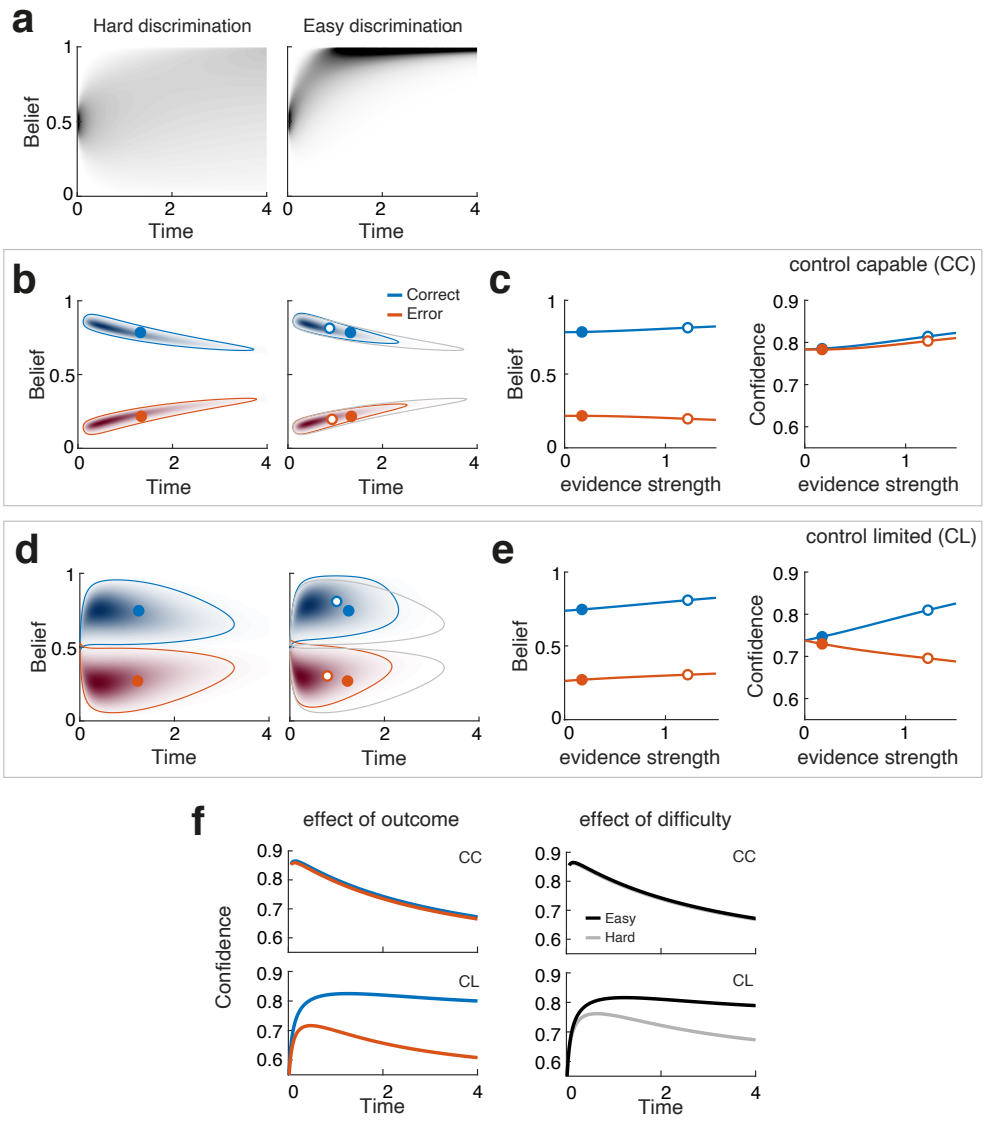


Fig. 5. Control limitations shape decision confidence. (a) Temporal evolution of belief for a difficult (left) and easy (right) stimulus conditions. The probability distribution of beliefs (Methods) is normalized at each time. **(b)** Left. Joint distribution of belief (at decision time) and RT for correct and error trials for the difficult condition (a, left) for a control capable agent ($\beta = 2^8$). Circles represent the mean. Line demarcates the region enclosing 90% of the probability mass. Right. Same but for the easy condition in (a, right). The means and regions of high probability from the hard condition are also shown for comparison. **(c)** Left. Mean belief at decision time as a function of the strength of evidence for correct and error trials. Circles show the values used in (a,b). Right. Decision confidence (equal to 1-g if $g < 0.5$) as a function of the strength of evidence. **(d,e)** Same as (b,c) but for a moderately control-limited agent ($\beta = 2^{3.5}$). **(f)** DC conditioned on RT as a function of outcome (left) and difficulty (right) for control-capable (CC; top) and control-limited (CL: bottom) agents.

asymmetric with respect of outcome (Fig. 5d, right). Intuitively, this occurs because the control-limited policy, despite being outcome-symmetric, is less restrictive in terms of the values of belief and RT where commitment is possible, so when beliefs are strongly biased by the evidence, DC ends up also reflecting this bias (Fig. 5d, right, Fig. 5e, left). Since errors occur when the agent

believes L is the correct choice ($g < 0.5$), the biasing of these beliefs by the evidence (towards R, i.e., towards $g = 1$) implies a more undecided state (g closer to 0.5), i.e., lower DC. This process results in opposing trends for DC as a function of the strength of evidence for correct and error trials (Fig 5e, right), demonstrating that optimal control-limited policies possess good CR. CR is generally observed in psychophysical experiments[61,63,67–71], but had so far been unaccounted within a normative sequential sampling framework (see Discussion).

Control-capable agents have poor CR because, when choices are triggered by hard bounds on belief, the only quantity that can shape decision confidence is RT (through the time-dependence of the decision bounds). In fact, if the bounds were constant, as in the SPRT, decision confidence would be identical for all choices, as noted early on[72]. Thus, for agents using these kinds of policies, decision confidence conditional on RT is independent of any other aspect of the problem, such as trial outcome or strength of evidence (Fig. 5f, top). In contrast, decision confidence in control-limited agents is shaped by all factors that affect the beliefs of the agent before commitment, and is therefore larger for correct than error trials, and for easier compared to than harder conditions (Fig. 5f, bottom). In sum, the stochastic nature of commitment imposed by the control limitations of an agent provides a natural explanation for the coupling between DC and the underlying factors that shape the beliefs of an agent during a perceptual decision.

Realization of the control-limited regime

In addition to CR (Fig. 6a, left), control-limitations have signatures at the level of RT and also at the level of accuracy conditional on RT (i.e., time-dependent accuracy or TDA[73], shown here averaged across difficulties). Errors following the control-limited policy are faster than correct choices (Fig. 6a, middle), as already evident in Fig. 5d. Errors tend to occur earlier because, on the one hand, the stochastic control-limited policy allows commitment with ambivalent beliefs and, on the other hand, these beliefs are more likely earlier on, as the belief of the agent aligns with the evidence as the trial progresses (Fig. 5a,d. Supplementary Fig. 2). In addition, TDA has a characteristic profile of initial growth and, if the control ability of the agent is low, it saturates to a roughly constant value (Fig. 6a, right). Qualitatively, these features are robust for control-limited agents, regardless of the specific β and λ of the agent (Fig. 6b). This is in contrast to the behavior of the FAB agent (or agents with very large β). As we showed in Fig. 3f, unless the stakes of the task are enormous, error RTs are longer than those of correct trials (although by a small amount. The hard bounds on accumulated evidence of the FAB policy decay relatively slowly, and this leads to a small outcome dependence of RT – Supplementary Fig. 1) and to a monotonically decaying TDA[48].

These data show that it should be possible to identify control-limited behavior based on these features. A difficulty, however, is that it is not trivial to manipulate experimentally the control-ability of a subject performing a discrimination task. Particularly in the case of human subjects, given instructions to perform, subjects will generally mobilize cognitive resources to comply. We reasoned that a strategy to find the control-limited regime would be to focus on situations where there is little incentive to invest resources in the task. First, we note that although so far we've treated the control ability β as a property of the subject, the ability to exercise control is both a dynamic and a limited resource[30,74–76]. Thus, it is expected that subjects will shape their allocation of control taking into account the gains they might experience from different allocation policies[32,77].

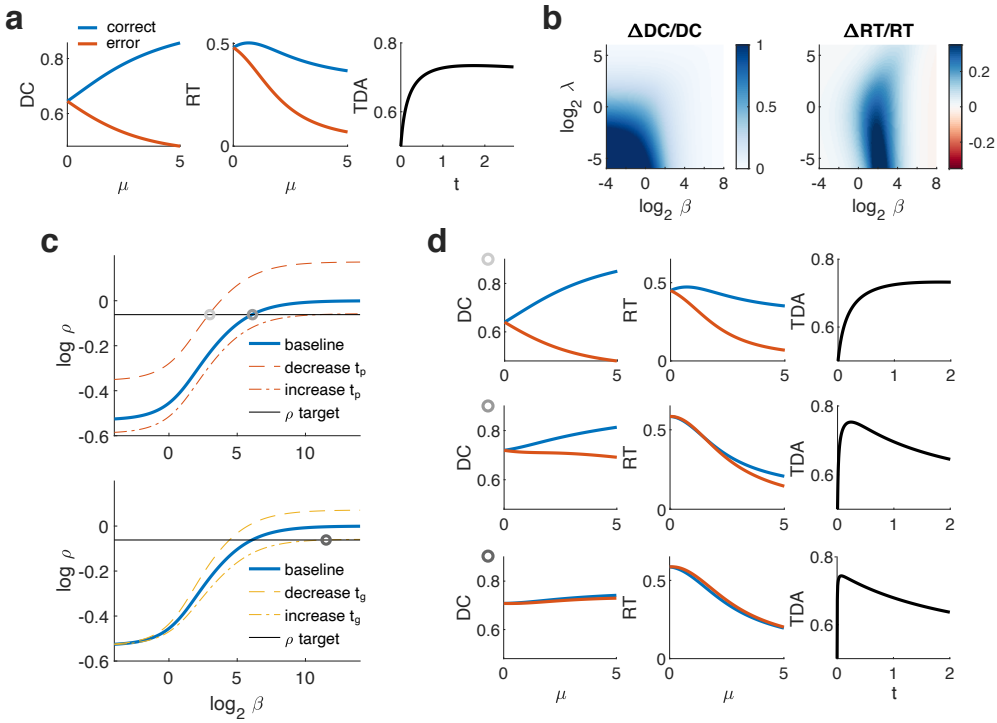


Fig. 6. Identifying the control-limited regime. **(a)** Example of outcome-dependence of DC and RT as a function of difficulty (showing clear CR and fast errors) and TDA, in the control-limited regime. **(b)** Magnitude of ΔDC/DC (which is equal to DC_{corr} - DC_{err} averaged across difficulties, relative to the same average regardless of outcome) and ΔRT/RT (same definition but for RT) in the space of β and λ. **(c)** Reward rate ρ as a function of the control ability β. Top and bottom show the same baseline situation, together with an estimated constant target reward rate, and the modified reward rate profiles under a change in the time penalty for errors t_p (top; equivalent to a change in time-pressure) and the time-scale of inference t_g (bottom; equivalent to a change in difficulty). **(d)** DC, RT and TDA (as in (a)) for the baseline situation (middle; signalled by a gray circle in (c, top)), a situation with a lower value of β that keeps the same target ρ under an increase in time-pressure (top; signalled by a light-gray circle in (c, top)), and a situation with a higher value of β that keeps the same target ρ under an increase in difficulty (bottom; signalled by a dark-gray circle in (c, bottom)).

As we showed in Fig. 4, the reward rate of the agent increases monotonically with β. Thus, in principle, a reward-maximizing agent should seek to increase the amount of invested control. In practice, however, subjects are expected to use satisficing, rather than maximizing strategies[78]. Alternatively, the marginal utility of reward (rate) is expected to decrease when the agent is satisfied[79,80].

We thus consider a realistic setting where an agent is performing a task whose parameters have been set so that the agent is close to the point of satisfaction using a certain amount of control (Fig. 6c). How is the agent expected to re-allocate control under different task manipulations? We considered two manipulations that are commonly used: changing time-pressure and changing the overall difficulty of the discrimination task. In our model, the time pressure is effectively controlled by the task stakes S (Figs. 3,4). Low stakes normatively induce time-pressure by effectively penalizing long RTs in terms of reward rate (Fig. 3). In practice, we lower the stakes by lowering the penalty time t_p after an error. Difficulty is controlled by the inference time-scale t_g. The lower t_g, the shorter the time that it takes to identify the latent state on average, i.e., the easier the task. Decreasing

the error penalty or decreasing the difficulty will both increase the reward rate of the agent, but if the agent is already close to its target reward rate, then the agent can stay at the target by using a lower β (Figs. 6c top, 6d bottom). Conversely, in response to the opposite manipulations, the agent should invest more control to stay at the target (Figs. 6c bottom, 6d top). Lower difficulties and time pressure are thus associated with control-limited phenotypes, whereas emphasis on accuracy and difficult discriminations correspond to control-capable behavior (Fig. 6d) These considerations suggest that control-limited behavior might be observed in conditions with high time pressure and easy discriminations. Indeed, this is in agreement with a substantial body of work in human decision making. Errors tend to be faster than correct trials under speed emphasis and in easy tasks[8,69,81,82], but slower than corrects under accuracy emphasis and hard discriminations[8,53,69,72,82]. CR has also been described to increase with time pressure[69,70]. As far as we can tell, the effect of manipulations of difficulty or time-pressure on the TDA has not been quantified in human decision-making, and thus remains an untested prediction of our theory.

Previous studies have focussed on post-decisional processing as a mechanism for producing CR in a sequential sampling framework[69–71,83]. In this kind of models, choices are still triggered when the accumulated evidence hits a bound (hence, choice and RT phenomenology is not affected), but DC depends on evidence accumulation after a decision is made. Typically, confidence is assumed to be a function of the value of the decision variable at some fixed time after commitment, referred to as the inter-judgement time. Post-decisional processing naturally leads to CR[16,69], and previous work has shown that CR indeed grows when the inter-judgement time is experimentally increased[71,83]. Thus, both post-decisional and control-limitations produce robust CR in a sequential sampling setting. The two mechanisms, however, are clearly distinct. A FAB agent using post-decisional processing for DC will have the same outcome of RT and TDA curve as the standard FAB agent. Thus, these behavioral signatures can be used to distinguish control-limitations from post-decisional processing.

Control limitations and decision lapses

During sensory discrimination experiments, lapses are identified by a saturation of the psychometric function to a value different from one or zero, signaling errors that don't have a sensory origin. Since action-selection under the default policies we consider is stimulus-independent, it is expected that control-limited agents will lapse if their control ability is sufficiently low.

Indeed, the psychometric function of the control-limited agent starts showing lapses as β decreases (Fig. 7a,c). Lapses appear when the probability of committing to either option is still close to the default rate $\lambda/2$ even under complete certainty about the sign of the latent variable, i.e., for beliefs $g = 0, 1$ (Fig. 7b). Avoiding lapses requires being able to form the appropriate beliefs based on sensory evidence, and being able to act on the basis of those beliefs. The control-limited agents we model are capable of the former process, but may not be capable of the latter. As the control ability of the agent increases, response probability becomes more strongly dependent on its beliefs. In particular, the probability of a correct response under sensory certainty becomes large, and the corresponding error probability becomes zero, and lapses disappear (Fig. 7b).

In most psychophysical experiments, one uses behavior to infer the sensory limitations of a subject, for instance through the slope of the psychometric function at the categorization boundary.

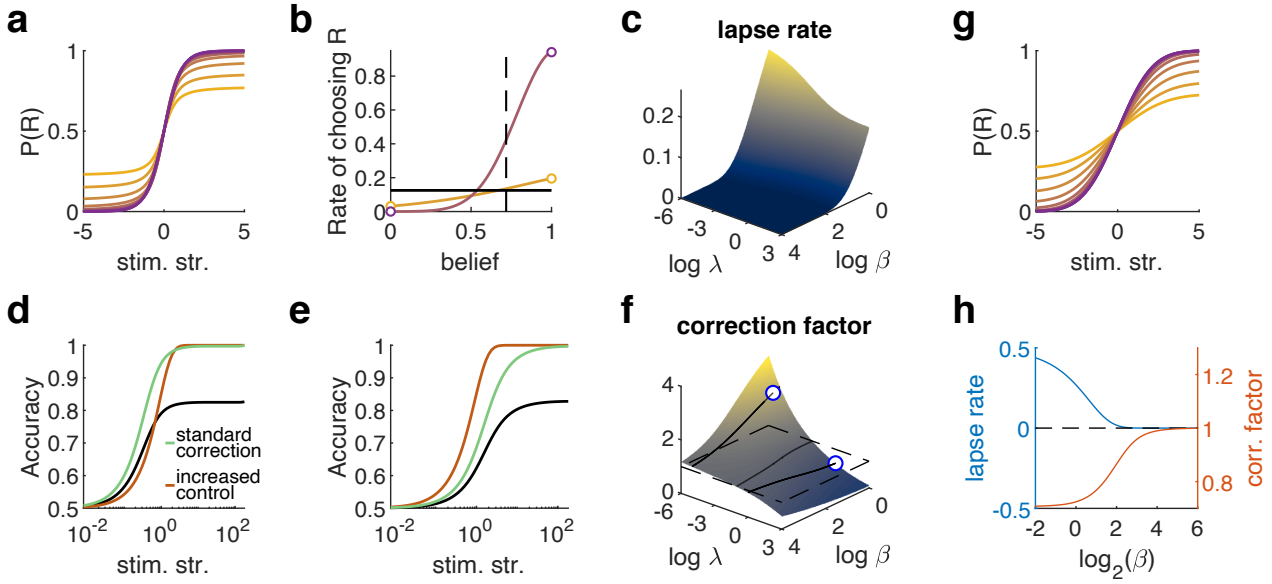


Fig. 7. Handling decision lapses produced by control limitations. **(a)** Psychometric functions display increasing lapse rates as the control-limitations of the agent grow (orange to blue). **(b)** Decision policies as a function of belief (at time $t = 0$) for two of the agents in (a) (same color scheme). The black horizontal line represents the default stimulus-independent policy. **(c)** Lapse rate as a function of the control ability β and default response rate λ. **(d)** Psychometric functions for a speed-biased control-limited agent (black), its standard correction (green; obtained by simply scaling the black curve until its asymptote reaches 1), and the correction obtained by setting $\beta^{-1} = 0$ (red; which is the psychometric function of the FAB agent). **(e)** Same as (d) but for an accuracy-biased agent. **(f)** Ratio between the slope of the psychometric function corrected with the standard method and the psychometric function of the FAB agent, as a function of β and λ. At the dashed plane both corrections agree. The two circles show the examples in (d) and (e). **(g)** Psychometric functions in a 'single sensory sample' model (equivalent to a SDT setting) as a function of the control ability of the agent. **(h)** Lapse rate (blue) and ratio between the slopes of the two corrected psychometric functions (same as (f)) for this model.

Because lapses change the shape of the psychometric function, they obscure the true psychophysical abilities of a subject. Developing methods to recover sensory limitations in the presence of other, non-sensory, processes shaping behavior, has been a critical problem in the history of psychophysics which, for instance, gave rise to the development of Signal Detection Theory[84] (SDT).

The standard approach to recover a 'clean' estimate of stimulus discriminability in the presence of lapses is simply to scale up the slope of the psychometric function until its asymptotes reaches one and zero[85]. This is appropriate when lapses reflect inattention[86]. On the other hand, when lapses result from control limitations, the proper approach to recover the true sensory limitations of the agent is to examine their psychometric function at $\beta^{-1} = 0$, i.e., in the absence of any limitations in control. To compare the performance of both approaches, we considered the result of applying both types of corrections to the psychometric function of a control-limited agent. Interestingly, the two approaches don't, in general, coincide (Fig. 7d-f). In fact, the 'cleaned' psychometric slope obtained using the standard approach can either over- or under-estimate the slope of the psychometric function of the FAB agent, depending mainly on the default response rate λ (Fig. 7f). Consider the case of an agent whose control-limitations produce a significant lapse rate (Fig.

7d-e). Small values of λ describe situations where the default time to respond is long compared to time-scale t_g of the inference process. In this case, the default policy of the agent overemphasizes accuracy over speed, leading to a steeper psychometric function compared to the FAB agent (at the cost of a lower reward rate). In these conditions, the standard correction overestimates the optimal psychometric slope (Fig. 7d). In contrast, when λ is large, the suboptimal reward rate of the control-limited agent comes from overemphasis of speed over accuracy, and this is associated to an underestimation of the optimal psychometric slope by the standard correction (Fig. 7e).

The discrepancy between the two correction methods is still there when one considers a simpler decision problem where an agent only receives one sample of sensory evidence (equivalent to a SDT setting). Here, one can also define optimal control-limited policies which will display lapses if the control ability of the subject is sufficiently low (Fig 7g; such policies are mathematically equivalent to those in Pisupati et al.[86], but see Discussion). In this setting, the pure 'sensory-limited' psychometric function does not depend on speed-accuracy considerations, and the standard correction factor always underestimates the true psychometric slope (Fig. 7h). Furthermore, comparing the lapse rate of the agent and the correction factor as the control ability of the agent grows, it is apparent that the correction factor still underestimates by the time the lapse rate reaches zero (a feature also present in the sequential decision problem, i.e., compare Fig. 7c and Fig. 7f at $\beta \sim 2^{2.5}$). This implies that saturation of the psychometric function to 1 or 0 does not automatically guarantee that the psychometric slope reflects the agent's true sensory limitations (Discussion).

In summary, control limitations naturally lead to lapses in decision-making. Our results show that the proper correction to the observed psychometric function depends on how lapses are generated, that the standard correction is in general not correct when lapses are due to control limitations, and that corrections might still be needed even if lapses are not fully apparent.

Sequential dependencies and decision biases

In laboratory settings, where many decisions are performed during an experiment, it is often observed that behavior in one trial can be partly explained by events taking place in past trials[26,27,87–89]. Many different forms of such sequential dependencies have been described, reflecting different processes including, for instance, reinforcement learning[88] or bayesian inference[87]. Whereas sequential dependencies are often (but not always[87]) maladaptive within the short term context of the task, they are typically adaptive when one considers longer-term environmental regularities. Such situations, where the short-term context and the long-term environment have opposing demands, are exactly the ones benefitting from control, which suggests that control-limitations might provide a natural framework for describing some forms of sequential dependency.

We consider three classes of sequential dependencies, grouped according to the quantity that is updated from one trial to the next. One class corresponds to updating the predisposition of choosing an action before the stimulus is observed, which can be modelled using biased default response policies (Fig. 8a-c). Another class corresponds to updating the value of the different actions, depending on previous events (Fig. 8d-f). A third class corresponds to updating the prior belief of the agent about the upcoming stimulus caterory (Fig. 8g-i). We devised a procedure for deriving optimal policies that incorporate each of these three forms of trial-to-trial updating (Supplementary Information). Importantly, each class can be used to describe a number of qualitatively different

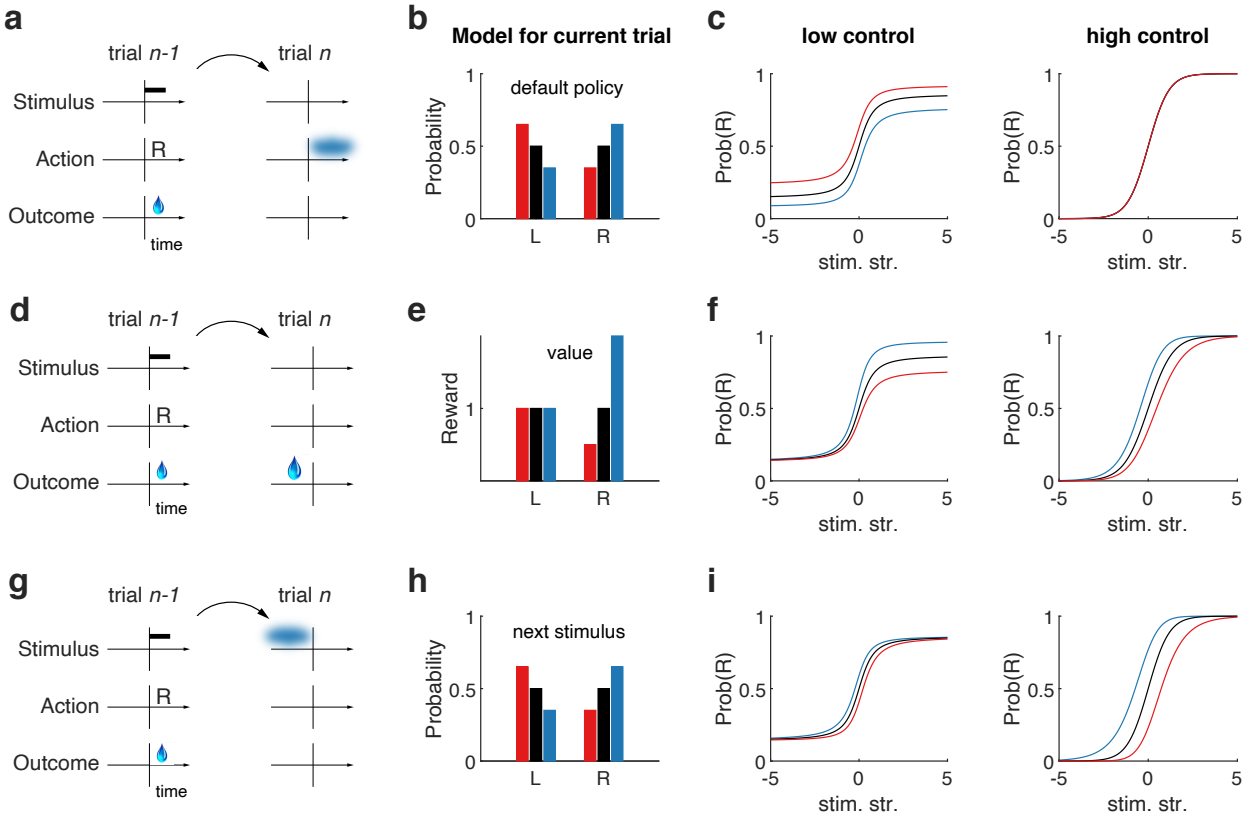


Fig. 8. Policies incorporating sequential dependencies. **(a)** Schematic description of a class of sequential dependencies where previous trials lead to a bias in the default action policy, depicted as a shading over one of the actions in the current trial **(b)** Three scenarios where the default policy is unbiased (black) or biased towards action R (blue) or L (red; Methods). **(c)** Psychometric functions for each of the three cases in (b) for control-limited (left) and control-capable (right) agents. **(d)** Similar to (a), but where trial history shapes the current value of the action that the agent just performed. **(e)** Three scenarios where the agent choose R in the previous trial and the history of choice-outcomes experienced leads to the reward magnitude in the current trial being modelled as higher (blue), lower (red) or average (black). **(f)** Same as (c) for value biases. **(g)** Similar to (a) but where trial history shapes the agent's prior beliefs about the upcoming stimulus, depicted as a shading over one of the stimulus categories before the vertical bar marking stimulus onset. **(h)** Three scenarios where the agent believes the latent variable in the current trial is more likely to be positive (blue), negative (red) or is unbiased (black). **(i)** Same as (c) but for biases in stimulus probability.

sources of sequential dependence. For instance, updates in the default probability of choosing an action can depend on the previous choice, or on an interaction between the previous choice and outcome. Our grouping into classes thus reflects the target of the cross-trial updating, not the events that cause the update.

To reveal the effects of these different forms of sequential dependence, we plot the psychometric functions of the agent conditioned on the relevant event in the previous trial. We use as a baseline an agent whose control limitations lead to a substantial lapse rate, and then show how the effect of each form of sequential dependence on the psychometric function is modified as the agent becomes control-capable. This strategy helps evaluate how control limitations shape the pattern of sequential dependencies in each class. The signature of sequentially updating the bias of the default policy, is

a symmetric vertical displacement in the psychometric function, as observed, for instance, by Scott et al.[90] (Fig. 8c, left). This is because increasing the probability of one action automatically implies lowering the probability of the other. These vertical shifts are unoccluded when the psychometric function does not saturate to one or zero, which is the case if the agent is sufficiently control limited to show lapses. Because sequential biases reflect the default action policy, they disappear under conditions of high control (Fig. 8c, right).

Agents might, on the other hand, use their history of successes and failures to sequentially update the value of each action (Fig. 8d), instead of the default probability of choosing it. In experiments where rewards and penalties are fixed, this would be a form of suboptimality, but it might be expected if subjects have the wrong model for the task, and incorrectly attribute fluctuations in average value across trials (due to variable proportions of incorrect choices) to fluctuations in the single-trial value of an action[91]. Because only the value of the action that was just produced is updated (Fig. 8e), this type of sequential dependencies lead to asymmetric modulations of the psychometric function, in which the amount of bias is proportional to the likelihood of repeating the action (Fig. 8f, left), as was recently observed[86]. In this case, although the sequential bias is still there for control capable agents, the marked asymmetry is almost completely eliminated (Fig. 8e, right), because the probability of repeating the action is already saturated at its maximum value of 1 when lapses disappear.

A final scenario we consider is an update in the prior belief of the agent about the stimulus (Fig. 8g-h) which might arise, for instance, if there are across-trial correlations in the value of the latent variable which the agent is learning[89]. Typically, updating of stimulus priors is expected to lead to horizontal displacements in the psychometric function (Fig. 8i, left). For a given magnitude of the bias in probability (Figs. 8b,h), the changes to the psychometric function are smaller when the updated probabilities refer to the stimulus prior compared to the case when they reflect the passive action policy (compare Figs. 8c,i left). This is because the behavior of control limited agents is only weakly adapted to the task demands and reflects to a large extent their passive policies. Thus, because for this class of sequential dependencies, choice biases are adaptive, they grow with the control abilities of the agent (Fig. 8i, right). We conclude that the shape of the modifications in the psychometric function due to trial history, together with their dependency on the agent's control ability, can be used to infer which aspect of the decision-making process is being updated across trials.

Discussion

We have systematically characterized how to optimally trade-off control and performance costs in perceptual decision making. We have considered stimulus-independent default policies to highlight the need for control in order to achieve good performance. Our default policies were also stochastic, as a means of phenomenologically describing all task-independent influences that might result in specific choices made at specific times. This type of default behavior results in optimal policies that have the form of smooth decision bounds. This means that there is no deterministic decision rule specifying when commitment will happen. Instead, accumulated evidence controls the probability of commitment, which transitions in general from a zero (or low) state, to a high state as the evidence favors more clearly one of the options (Fig. 2).

Our model is an extension of the KL control framework[34–37]. Initially developed as providing efficient solutions to optimal control problems in engineering, this framework is also finding applications in behavioral neuroscience, such as tractable solutions to the flexible-replanning problem[38], or as an explanatory framework for capacity limitations in action-policy complexity[36,39]. We have shown that KL control is also useful to study a variety of topics in sensory-limited sequential decision problems. From a methodological perspective, our main contribution is a flexible mathematical framework for optimal control of non-stationary point processes (Supplementary Information). Here, these model the timing of discrete actions, but the framework can be naturally extended to model spiking neural activity, an approach we are currently pursuing.

Although our decision-making model is not mechanistic, in principle, an approximation of the control-limited policies we have found is compatible with a standard deterministic decision bound if one assumes that the true decision variable is a weighted sum of the belief-dependent decision variable $x(g,t)$ we have described (Supplementary Fig. 1; Supplementary Information), and a stimulus-independent stochastic term, which would induce stochasticity in the transformation from $x(g,t)$ to action (in a recent study, this stimulus-independent term has been proposed to carry information about the timing of the onset of the stimulus[92]). The relative weight of these two components of the decision variable would be determined by the control ability of the agent. Control-capable agents would be able to suppress task-independent sources of input to the decision variable of the problem. A stochastic additive contribution to the decision variable can also be qualitatively approximated by trial-to-trial variability of a deterministic within-trial decision bound. This form of trial-to-trial variability has been considered in the past[8], and gives rise to phenomenology which is partially overlapping with that evident in agents using optimal control-limited policies. Our results can thus be interpreted as providing a normative grounding for this type of trial-to-trial variability in terms of control limitations.

Our framework assumes that agents seek to control a default exploratory policy. What exactly is the relationship between exploration and control? In the context of the exploration-exploitation trade-off, which is usually studied using n-armed bandit tasks[43,93], some studies have suggested that, in fact, it is exploitation that should be considered as an automatic default, and that exploratory choices require cognitive control[93]. For instance, behavior driven by Pavlovian associations might be construed as automatic. However, other studies have found that, even in bandit tasks, cognitive load (which limits the ability to use cognitive control) makes behavior more stochastic[39,94]. A recent study directly probed the relationship between cognitive load and directed versus random exploration, and found that only directed exploration is down-regulated in high-load conditions[95]. Besides, it is critical to note that exploitative behavior in an n-armed bandit task is very different from what might constitute exploitative behavior in a perceptual decision-making task, which requires temporal accumulation of evidence. Evidence accumulation is a form of working memory, expected to require cognitive control[96].

Control-limitations robustly shape the phenomenology of decision making. One consequence of making decisions using probabilistic decision bounds is that it automatically results in good confidence resolution (CR; Figs. 5-6). CR arises naturally in normative models of decision making based on signal detection theory[63,64,97] (SDT). But these models – which correspond in a sequential sampling setting to the use of vertical decision bounds – are clearly suboptimal when sensory evidence arrives in time, and are unable to account for the speed accuracy trade-off. However, at the

cost of giving up on explaining RT, models with vertical decision bounds allow the decision variable at the moment of commitment to be sensitive to the way in which sensory evidence shapes the belief distribution, which is fundamentally what CR necessitates (Fig. 5). On the other hand, somewhat counterintuitively, the fully control-capable FAB agent of the sequential sampling framework has poor CR[17] (Figs 3,5). Although being more confident in one's knowledge when it is in fact correct seems advantageous, specially in a social setting[98], it turns out that it is not optimal from the point of view of maximizing performance. In the absence of control limitations, choices should be made only based on instantaneous belief and elapsed time[15,62], and thus outcome can only affect DC through its effect on RT (Figs. 3,5).

Given the widespread empirical observation of CR[60,67,72], decision theorists have sought ways of obtaining robust CR within a sequential sampling setting. The standard solution relies on post-decisional processing of DC[61,69–71,83]. Separating in time decision commitment and DC permits using (effectively) horizontal bounds for choice while keeping vertical bounds (*a la* STD) for computing DC, which produces both speed-accuracy trade-offs and CR. CR has been shown to vary when the window of post-decisional integration is causally manipulated[71,83], and at the same time lower CR is observed when choice and decision confidence are reported simultaneously by design[62,65], suggesting that post-decisional integration does contribute to observed CR. Is post-decisional processing adaptive? There is conflicting evidence on this issue. Some work has pointed out to a need for post-decisional time to explicitly compute DC under some conditions[61], and there are suggestions that some frontal areas as specifically involved in the computation of DC. At the same time, Bayesian confidence is an instantaneous function of accumulated evidence and elapsed time[15,48,62], confidence and choice can be reported simultaneously[62,65,71], and both choice and confidence-related signals have been observed in the same parietal circuits[99–101].

Although control-limited policies and post-decisional processing both produce CR robustly, they are clearly distinct, and result in opposing trends in terms of the outcome-dependence of RT and of the TDA profile, quantities which are unmodified by post-decisional processing. Unless the stakes of the task are extremely high, the outcome-dependence of RT reverses sign as a function of β (Figs 3, 6). Such sign-reversal is well documented empirically: errors tend to be faster than correct trials in task settings which encourage speeded responding and where discriminations are easy[8,69,81,82], whereas when the task emphasizes accuracy and for more difficult discriminations, it is correct trials that tend to have shorter RTs[8,53,69,72,82]. Our results suggest a normative explanation for this organization in terms of the connection between motivation and control: faced with the choice between continuing to invest control to increase reward with little marginal utility, and investing less control with little loss in satisfaction, agents will choose the latter (Fig. 6c). In fact, in our formalism, control and motivation are tightly related. The control ability of the agent can only be defined relative to the reward available from the different actions (i.e., β^{-1} has units of reward – Supplementary Information). Motivation has been construed as a map that specifies the utility of outcomes[102], which is posited to vary across different internal states. Control-limitations provide another dimension that affects the efficiency with which outcomes can mobilize behavior. The same reward will be differently effective in driving adaptive responding, even under equal internal states, depending on how such adaptive responses relate to an agent's default policies and its capability for cognitive control.

We have shown that there is a natural connection between decision lapses and the ability to

control task-independent default policies (Fig. 7a-c). Lapses arising from control limitations are formally similar to lapses construed as a form of exploration, as recently suggested by Pisupati et al[86]. However, whereas we view lapses as essentially reflecting a limitation of the agent, Pisupati et al. construe them as adaptive in perceptual decision-making tasks – because the agent will perceive action-outcome as being stochastic due to errors caused by their sensory limitations. We think it's unlikely that the existence of sensory errors *per se* will generally be modelled by subjects, including rodents, as reflecting probabilistic reward contingencies, given that rats in some difficult perceptual decision-making tasks do not lapse[3]. However, exploratory strategies would be adaptive if the agent feels like it has not yet learned the correct model of the environment[43], particularly at early stages of training. While it is possible that in some experiments, rodents (incorrectly) model the environment as perpetually changing – and thus sustain a stationary lapsing policy – we suggest that it might be reasonable to interpret this as a limitation to suppress a default tendency towards exploration. Independently of their interpretation, lapses obscure the true psychophysical performance of the agent. We suggested that the necessary correction when lapses reflect control limitations (or exploration) is in general different from the standard correction used in psychophysics[85,86]. The relationship between the two corrections depends on the speed, or accuracy emphasis imposed by the agent's default policy (Fig. 7d-f), but the two are still different even for SDT models (Fig. 7g,h). In fact, our results show that the slope of the psychometric function will not, in general, reflect the true sensory abilities of the subject even if the psychometric function saturates to one or zero (Fig. 7f,h). This implies that accurate assessments of sensory sensitivity require paying attention to the functional form of the psychometric function, i.e., not all sigmoidal functions are equally appropriate.

Recent studies[103] have used hidden Markov models to identify transitions between states characterized by different levels of engagement, and shown that this phenomenology accurately describes some types of decision lapses and some forms of sequential dependencies. In our study, we also considered sequential dependencies, but focused on whether it would be possible to identify the targets of cross-trial modification based on the relationships they induce between the psychometric functions calculated in successive trials (Fig. 8c,f,i). We showed that sequential changes in action priors, stimulus priors, or action-values are dissociable, specially for control-limited agents. Interestingly, the three corresponding patterns of change in the psychometric function have all been observed experimentally in different tasks[86,89,90]. In principle, our formalism could describe also cross-trial changes in control (Supplementary Information), which would correspond more directly to the approach of Ashwood et al.[103]. However, an adequate description of this phenomenology would require a normative framework for the optimal moment-to-moment allocation of control.

In behavioral economics, the dynamics of control as a limited resource is known as ego depletion[30,74–76]. The main finding is that subjects perform worse in a task requiring cognitive control after having participated in a previous cognitively demanding task (compared to controls). It is controversial whether the origin of this limitation has a computational origin[77,104] or whether it is the consequence of scarcity of a physical resource[30,104–106], but regardless of its mechanistic origin, an agent that is aware of this limitation should attempt to allocate the control expenditure in an advantageous way, in essence solving the hierarchical problem of optimizing task performance and control allocation simultaneously[32]. A theoretical framework describing this hierarchical control problem would be useful to provide a normative understanding of the dynamics of engagement in behavioral tasks, the causal triggers of state transitions, or even the discrete nature of extended

behavioral states.

Code Availability

Custom MATLAB scripts used to implement the mathematical framework and produce the figures are available upon request.

Acknowledgements

We thank Pietro Vertecchi, Tiago Costa, and Gautam Agarwal for discussions. J.C. was supported by a doctoral fellowship from the Fundação para a Ciência e a Tecnologia (FCT). AR was supported by the Champalimaud Foundation, a Marie Curie Career Integration Grant PCIG11-GA-2012-322339, the HFSP Young Investigator Award RGY0089, the EU FP7 grant ICT-2011-9-600925 (NeuroSeeker), and grants LISBOA-01-0145-FEDER-032077 and PTDC/MED-NEU/4584/2021 from the FCT.

Author contributions

J.C and A.R. conceived the project and the theory. J.C developed the theory and conducted the analysis. A.R. wrote the manuscript with feedback from all authors.

Competing Interests

All authors declare no competing interests.

References

1. Kiani, R., Hanks, T. D. & Shadlen, M. N. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience* **28,** 3017–3029 (2008).
2. Brunton, B. W., Botvinick, M. M. & Brody, C. D. Rats and humans can optimally accumulate evidence for decision-making. *Science* **340,** 95–98 (2013).
3. Pardo-Vazquez, J. L. *et al.* The mechanistic foundation of Weber's law. *Nature neuroscience* **22,** 1493–1502 (2019).
4. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review* **113,** 700 (2006).
5. Forstmann, B. U., Ratcliff, R. & Wagenmakers, E.-J. Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology* **67** (2016).
6. Link, S. & Heath, R. A sequential theory of psychological discrimination. *Psychometrika* **40,** 77–105 (1975).
7. Ratcliff, R. A theory of memory retrieval. *Psychological review* **85,** 59 (1978).

8. Ratcliff, R. & Rouder, J. N. Modeling response times for two-choice decisions. *Psychological Science* **9,** 347–356 (1998).

9. Usher, M. & McClelland, J. L. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review* **108,** 550 (2001).

10. Palmer, J., Huk, A. C. & Shadlen, M. N. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of vision* **5,** 1–1 (2005).

11. Wald, A. Sequential tests of statistical hypotheses. *The annals of mathematical statistics* **16,** 117–186 (1945).

12. Wald, A. & Wolfowitz, J. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics,* 326–339 (1948).

13. Gold, J. I. & Shadlen, M. N. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36,** 299–308 (2002).

14. Huang, Y., Hanks, T., Shadlen, M., Friesen, A. L. & Rao, R. P. How prior probability influences decision making: A unifying probabilistic model. *Advances in neural information processing systems* **25,** 1268–1276 (2012).

15. Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N. & Pouget, A. The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience* **32,** 3612–3628 (2012).

16. Calder-Travis, J., Bogacz, R. & Yeung, N. Bayesian confidence for drift diffusion observers in dynamic stimuli tasks. *BioRxiv* (2020).

17. Khalvati, K., Kiani, R. & Rao, R. P. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature communications* **12,** 1–16 (2021).

18. Howard, R. A. Dynamic programming and markov processes. (1960).

19. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 1998).

20. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence* **101,** 99–134 (1998).

21. Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications* **10,** 1–7 (2019).

22. Shiffrin, R. M. & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review* **84,** 127 (1977).

23. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annual review of neuroscience* **24,** 167–202 (2001).

24. Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. Conflict monitoring and cognitive control. *Psychological review* **108,** 624 (2001).

25. Stroop, J. R. Studies of interference in serial verbal reactions. *Journal of experimental psychology* **18,** 643 (1935).

26. Yu, A. J. & Cohen, J. D. Sequential effects: Superstition or rational behavior? *Advances in neural information processing systems* **21,** 1873 (2008).

27. Fründ, I., Wichmann, F. A. & Macke, J. H. Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of vision* **14,** 9–9 (2014).

28. Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. Decision making and the avoidance of cognitive demand. *Journal of experimental psychology: general* **139,** 665 (2010).

29. Dixon, M. L. & Christoff, K. The decision to engage cognitive control is driven by expected reward-value: neural and behavioral evidence. *PloS one* **7,** e51637 (2012).

30. Baumeister, R. F., Vohs, K. D. & Tice, D. M. The strength model of self-control. *Current directions in psychological science* **16,** 351–355 (2007).

31. Smith, V. L. & Walker, J. M. Monetary rewards and decision cost in experimental economics. *Economic Inquiry* **31,** 245–261 (1993).

32. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79,** 217–240 (2013).

33. Lieder, F. & Griffiths, T. L. Strategy selection as rational metareasoning. *Psychological review* **124,** 762 (2017).

34. Kappen, H. J. Linear theory for control of nonlinear stochastic systems. *Physical review letters* **95,** 200201 (2005).

35. Todorov, E. Efficient computation of optimal actions. *Proceedings of the national academy of sciences* **106,** 11478–11483 (2009).

36. Tishby, N. & Polani, D. in *Perception-action cycle* 601–636 (Springer, 2011).

37. Ziebart, B. D. Modeling purposeful adaptive behavior with the principle of maximum causal entropy (2010).

38. Piray, P. & Daw, N. D. Linear reinforcement learning: Flexible reuse of computation in planning, grid fields, and cognitive control. *BioRxiv,* 856849 (2020).

39. Lai, L. & Gershman, S. J. in *Psychology of Learning and Motivation* 195–232 (Elsevier, 2021).

40. Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E. & Tishby, N. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789* (2015).

41. Cover, T. M. *Elements of information theory* (John Wiley & Sons, 1999).

42. Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909* (2018).

43. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition* **173,** 34–42 (2018).

44. Puterman, M. L. Markov decision processes. *Handbooks in operations research and management science* **2,** 331–434 (1990).

45. Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning* **22,** 159–195 (1996).

46. Rao, R. P. Decision making under uncertainty: a neural model based on partially observable markov decision processes. *Frontiers in computational neuroscience* **4,** 146 (2010).

47. Yu, X. & Lewis, E. R. Studies with spike initiators: linearization by noise allows continuous signal modulation in neural networks. *IEEE Transactions on Biomedical Engineering* **36,** 36–43 (1989).

48. Drugowitsch, J., Moreno-Bote, R. & Pouget, A. Relation between belief and performance in perceptual decision making. *PloS one* **9,** e96511 (2014).

49. Malhotra, G., Leslie, D. S., Ludwig, C. J. & Bogacz, R. Time-varying decision boundaries: insights from optimality analysis. *Psychonomic bulletin & review* **25,** 971–996 (2018).

50. Davidson, J. D. & El Hady, A. Foraging as an evidence accumulation process. *PLoS computational biology* **15,** e1007060 (2019).

51. Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion decision model: Current issues and history. *Trends in cognitive sciences* **20,** 260–281 (2016).

52. Ditterich, J. Stochastic models of decisions about motion direction: behavior and physiology. *Neural networks* **19,** 981–1012 (2006).

53. Swensson, R. G. The elusive tradeoff: Speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics* **12,** 16–32 (1972).

54. Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of neuroscience* **22,** 9475–9489 (2002).

55. Botvinick, M., Braver, T., *et al.* Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology* **66,** 83–113 (2015).

56. Lichtenstein, S., Fischhoff, B. & Phillips, L. D. in *Judgment under uncenainty: Heuristics and biases* (eds Kahnemann, D., Slovic, P. & Tversky, A.) 306–334 (Cambridge University Press, Cambridge, 1982).

57. Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian probability: From neural origins to behavior. *Neuron* **88,** 78–92 (2015).

58. Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367,** 1322–1337 (2012).

59. Liberman, V. & Tversky, A. On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin* **114,** 162 (1993).

60. Baranski, J. V. & Petrusic, W. M. The calibration and resolution of confidence in perceptual judgments. *Perception & psychophysics* **55,** 412–428 (1994).

61. Baranski, J. V. & Petrusic, W. M. Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance* **24,** 929 (1998).

62. Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84,** 1329–1342 (2014).

63. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90,** 499–506 (2016).

64. Fleming, S. M. & Daw, N. D. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological review* **124,** 91 (2017).

65. Van Den Berg, R. *et al.* A common mechanism underlies changes of mind about decisions and confidence. *Elife* **5,** e12192 (2016).

66. Rausch, M., Hellmann, S. & Zehetleitner, M. Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics* **80,** 134–154 (2018).

67. Johnson, D. M. Confidence and speed in the two-category judgment. *Archives of Psychology* **34,** 1–53 (1939).

68. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455,** 227–231 (2008).

69. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review* **117,** 864 (2010).

70. Moran, R., Teodorescu, A. R. & Usher, M. Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive psychology* **78,** 99–147 (2015).

71. Desender, K., Donner, T. H. & Verguts, T. Dynamic expressions of confidence within an evidence accumulation framework. *Cognition* **207,** 104522 (2021).

72. Vickers, D. *Decision processes in visual perception* (New York, NY: Academic Press, 1979).

73. Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E. & Shadlen, M. N. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *Journal of Neuroscience* **31,** 6339–6352 (2011).

74. Vohs, K. D. *et al.* in *Self-regulation and self-control* 45–77 (Routledge, 2018).

75. Muraven, M. & Baumeister, R. F. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological bulletin* **126,** 247 (2000).

76. Hagger, M. S., Wood, C., Stiff, C. & Chatzisarantis, N. L. Ego depletion and the strength model of self-control: a meta-analysis. *Psychological bulletin* **136,** 495 (2010).

77. Kool, W., Shenhav, A. & Botvinick, M. M. Cognitive control as cost-benefit decision making. (2017).

78. Simon, H. A. Rational choice and the structure of the environment. *Psychological review* **63** (1956).

79. Reinagel, P. Rational regulation of water-seeking effort in rodents. *Proceedings of the National Academy of Sciences* **118,** e2111742118 (2021).

80. Glimcher, P. W. & Fehr, E. *Neuroeconomics: Decision making and the brain* (Academic Press, 2013).

81. Swensson, R. G. & Edwards, W. Response strategies in a two-choice reaction task with a continuous cost for time. *Journal of Experimental Psychology* **88,** 67 (1971).

82. Townsend, J. T., Ashby, F. G., *et al. Stochastic modeling of elementary psychological processes* (CUP Archive, 1983).

83. Yu, S., Pleskac, T. J. & Zeigenfuse, M. D. Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General* **144,** 489 (2015).

84. Tanner Jr, W. P. & Swets, J. A. A decision-making theory of visual detection. *Psychological review* **61,** 401 (1954).

85. Wichmann, F. A. & Hill, N. J. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & psychophysics* **63,** 1293–1313 (2001).

86. Pisupati, S., Chartarifsky-Lynn, L., Khanal, A. & Churchland, A. K. Lapses in perceptual decisions reflect exploration. *Elife* **10,** e55490 (2021).

87. Ashourian, P. & Loewenstein, Y. Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS one* **6,** e19551 (2011).

88. Mendonça, A. G. *et al.* The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. *Nature communications* **11,** 1–15 (2020).

89. Hermoso-Mendizabal, A. *et al.* Response outcomes gate the impact of expectations on perceptual decisions. *Nature communications* **11,** 1–13 (2020).

90. Scott, B. B., Constantinople, C. M., Erlich, J. C., Tank, D. W. & Brody, C. D. Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. *Elife* **4,** e11308 (2015).

91. Lak, A. *et al.* Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. *Neuron* **105,** 700–711 (2020).

92. Hernández-Navarro, L., Hermoso-Mendizabal, A., Duque, D., de la Rocha, J. & Hyafil, A. Proactive and reactive accumulation-to-bound processes compete during perceptual decisions. *Nature Communications* **12,** 1–15 (2021).

93. Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441,** 876–879 (2006).

94. Collins, A. G. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience* **35,** 1024–1035 (2012).

95. Cogliati Dezza, I., Cleeremans, A. & Alexander, W. Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General* **148,** 977 (2019).

96. Braver, T. S., Gray, J. R., Burgess, G. C., *et al.* Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. *Variation in working memory* **75,** 106 (2007).

97. Hangya, B., Sanders, J. I. & Kepecs, A. A mathematical framework for statistical decision confidence. *Neural Computation* **28,** 1840–1858 (2016).

98. Madirolas, G. & de Polavieja, G. G. Improving collective estimations using resistance to social influence. *PLoS computational biology* **11,** e1004594 (2015).

99. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *science* **324,** 759–764 (2009).

100. Fetsch, C. R. *et al.* Focal optogenetic suppression in macaque area MT biases direction discrimination and decision confidence, but only transiently. *Elife* **7,** e36523 (2018).

101. Fetsch, C. R., Kiani, R. & Shadlen, M. N. *Predicting the accuracy of a decision: a neural mechanism of confidence* in *Cold Spring Harbor symposia on quantitative biology* **79** (2014), 185–197.

102. Niv, Y., Joel, D. & Dayan, P. A normative perspective on motivation. *Trends in cognitive sciences* **10,** 375–381 (2006).

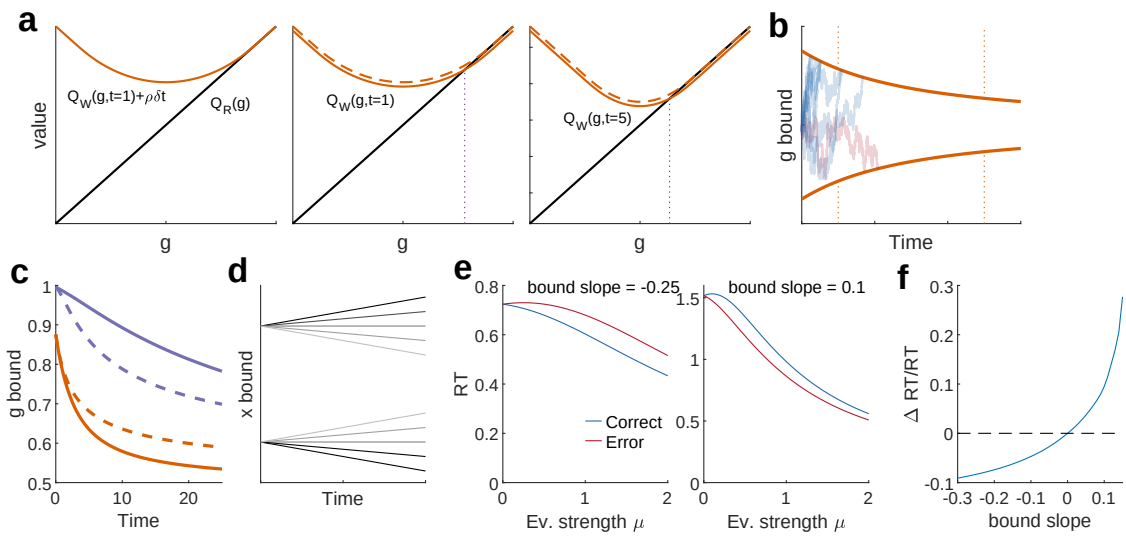
103. Ashwood, Z. C. *et al.* Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience* **25,** 201–212 (2022).

104. Shenhav, A. *et al.* Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience* **40,** 99–124 (2017).

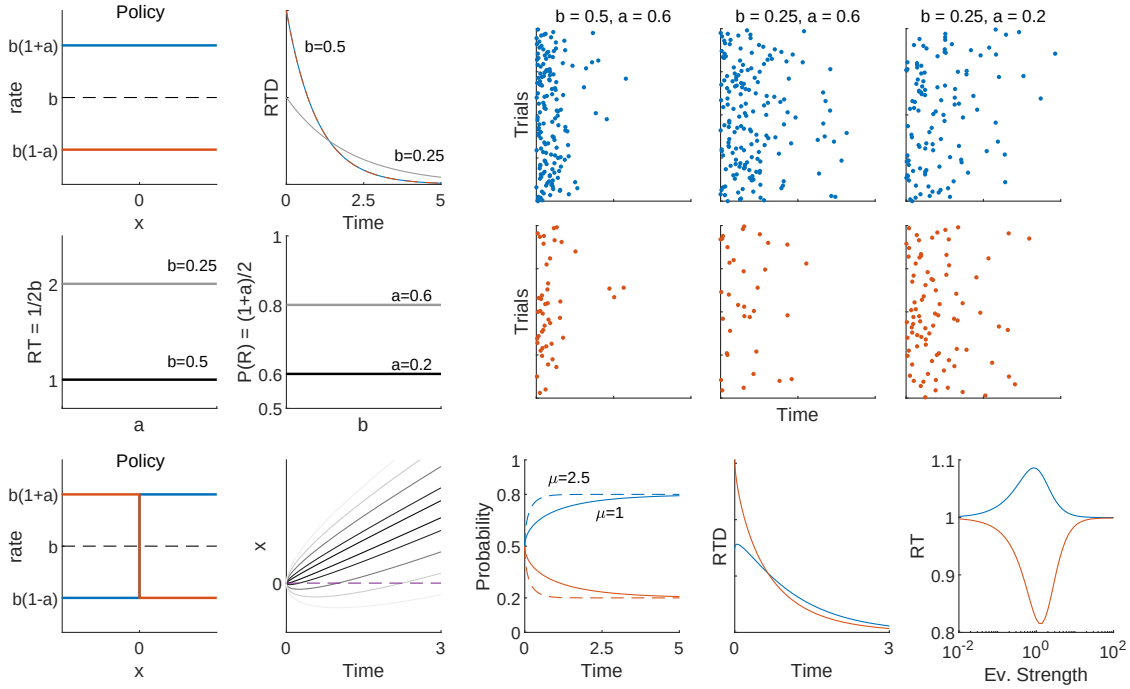
105. Fairclough, S. H. & Houston, K. A metabolic measure of mental effort. *Biological psychology* **66,** 177–190 (2004).

106. Gailliot, M. T. *et al.* Self-control relies on glucose as a limited energy source: willpower is more than a metaphor. *Journal of personality and social psychology* **92,** 325 (2007).

Supplementary Figures



Supplementary Fig 1. Construction of the optimal bounds for the FAB agent and relationship to the outcome difference in RTs. (a) Left: In black, action value of rightward choice as a function of belief g. In orange, action value of waiting for time $t = 1$ (a.u.), without considering the opportunity cost $\rho \cdot \delta t$. The orange curve is tangent to the black one, as the information gained by waiting cannot decrease the value. **Middle:** When considering now the opportunity cost of waiting (new solid line, previous curve is now dashed), the action values of choosing right and waiting intersect at the optimal decision bound in belief for that time (dotted line). **Right:** Same as middle but for a larger time value ($t = 5$), notice that the bound has decreased. **(b)** Optimal decision bounds on belief as a function of time. A few example trajectories are shown (blue for rightward choices, red for leftward choices). The values of time used in (a) are marked by the vertical dotted lines. **(c)** In solid lines, same decision bounds in belief as in Fig. 3a (excluding the lower value of the stakes). The orange bound is also the same shown in (b). In dashed lines, belief bounds that would be obtained if the bounds in the accumulated evidence x were constant and equal to their value at t=0. For the orange case, the dashed curve is above the solid one, suggesting that the real bounds in x are decreasing in time, while for the purple case, the dashed curve is below the solid one, suggesting that the real bounds in x are increasing in time. **(d)** In order to capture the qualitative difference in the RT conditioned on outcome ($\Delta RT/RT = (RT_{cor} - RT_{err})/RT_{total}$) observed in Fig. 3c, we considered a minimal model where the decision bounds in x are linear with variable slope and fixed intercept (i.e. same value at $t = 0$), such that the bounds can either be increasing or decreasing in time. **(e)** Two examples are shown for the mean RT of correct and error responses for different signs of the bound's slope. Left: example with negative slope, for which the RT of errors is larger than the RT for corrects. Right: example with positive slope, for which the RT of corrects is larger than the RT of errors. **(f)** plotting $\Delta RT/RT$ as a function of the bound slope reveals that a change in the sign of the latter is sufficient to induce a change in the sign of former, and that there is a one-to-one relationship between the two quantities. Interestingly, the relationship is highly non-linear.



Supplementary Fig 2. Properties of poisson rate policies. (a) Example of a constant rate policy (homogenous poisson race) as a function of the accumulated evidence x, with the parameter b determining the sum of the rightward (blue) and leftward rates (orange), and a determining the difference between them. **(b)** RTD produced by two constant rate policies with different values of b. The distributions are the same for rightward or leftward choices. **(c)** As a consequence, mean RTs only depend on b ($E[RT] = 1/2b$), as shown plotted as a function of a. **(d)** conversely, the proportion of rightward choices only depends on a ($P(R) = (1+a)/2$), as shown plotted as a function of b. **(e)** example trials visually demonstrating the results in (b-d). Top panels show rightward choices, bottom show leftward choices. In each column, the resulting RTDs are the same irrespective of choice. From the first to the second column, only b changes, so the choices overall become slower but the rightward proportion does not change. From the second to the third column, only a changes, so the RTDs stay the same but the rightward proportion changes. **(f)** Example of a simple monotonic rate policy: the rates are step functions of the accumulated evidence x, parametrized in the same way as before. As x changes sign, the rightward rate transitions from low to high, and the leftward rate does the opposite. The sum of both rates is still constant and only depends on b. **(g)** Distribution of x as a function of time for evidence strength $\mu = 1$. Several quantiles are shown. **(h)** As the mass of the distribution of x shifts towards positive values with time, the conditional probability of a rightward choice increases and the conditional probability of a leftward choice decreases. Increasing μ (solid vs dotted lines) changes the speed of convergence towards the asymptotic values, which are still $(1 \pm a)/2$. **(i)** This causes the RTDs for rightward and leftward choices to be different, as they are the product of the probability of making a choice at a given time (which still only depends on b and it is the same as in (b)) times the conditional probability of a given outcome (h). $\mu = 1$ in this example. **(j)** Mean RT conditioned on choice as a function of evidence strength (in log scale). The mean total RT is always the same because it only depends on b, but it differs conditioned on outcome as the evidence strength controls the shape of the curves in (h). Throughtout this example (h-j), $a = 0.6$ and $b = 0.5$.

Fig	Panel	Detail	t_g	t_p	t_i	λ	β	μ	ϕ	r	g_0
2	All		1	2^3	1	1			0.5	1	0.5
	a,b,c	CC:					2^8				
		CL:					2^2				
	b	$\lambda' = 1/4$									
	c							0.5			
4	All		1		1			-	0.5	1	0.5
	a-d	Solid blue		2^{-2}		2^{-5}					
		Solid gold		2^{-2}		2^3					
		Dashed blue		2^5		2^{-5}					
		Dashed gold		2^5		2^3					
	e-h	Blue				2^{-5}	2^4				
		Gold				2^3	2^4				
5	All		1	2^3	1	1			0.5	1	0.5
	a,b,d	Difficult						0.18			
	a,b,d	Easy						1.22			
	b,c,f	CC					2^8				
	d,e,f	CL					2^2				
6	All							-	0.5	1	0.5
	a		1	1	1	1	2^3				
	b		1	1	1						
	c	baseline	1	1	1	2					
		decreased t_p	1	$2^{0.7}$	$2^{0.7}$	2					
		increased t_p	1	$2^{1.1}$	$2^{1.1}$	2					
		decreased t_g	$2^{0.4}$	1	1	2					
		increased t_g	$2^{1.5}$	1	1	2					
		$\rho_{ideal} = 0.94\rho_{max}$									
		β_{ideal} values:									
		baseline					$2^{6.08}$				
		decreased t_p					$2^{2.91}$				
		increased t_g					$2^{11.43}$				
7	All		1	2^2	1			-	0.5	1	0.5
	a					2^{-3}	2^0-2^9				
	b	lower β				2^{-3}	$2^{0.5}$				
		higher β				2^{-3}	$2^{2.5}$				
	d	uncorrected				2^0	2^0				
	e	uncorrected				2^{-5}	$2^{0.5}$				
8	All		1	2^2	1	2^{-3}		-			
	b,c								$(0.65, 0.5, 0.35)$	1	0.5
	e,f								0.5	$(2, 1, 0.5)$	0.5
	h,i								0.5	1	$(0.65, 0.5, 0.35)$
	c,f,i	CL					$2^{0.5}$				
	c,f,i	CC					2^8				

Supplementary Table 1. Parameters values for all figures in the text.

Supplementary Note

1 Framework **34**
- 1.1 Markov Decision Processes: First exit formulation . . . 34
- 1.2 Kullback-Leibler control . . . 34
- 1.3 Equivalence between MDP and KL formulations . . . 36
- 1.4 Average-adjusted reward formulation . . . 38
- 1.5 Partial Observability: POMDPs . . . 41

2 Modeling a binary choice over a continuous latent state **41**
- 2.1 States and transitions . . . 41
- 2.2 Default policy . . . 42
- 2.3 Inference . . . 42
- 2.4 Dimensional analysis . . . 45
- 2.5 Bellman Equation . . . 45
  - 2.5.1 Initial and boundary conditions . . . 48
  - 2.5.2 Calculation of reward rate . . . 50
- 2.6 Optimal Policy . . . 50
- 2.7 Accuracy, reaction time and decision confidence . . . 51
  - 2.7.1 Decision lapses . . . 52

3 Numerical methods **52**
- 3.1 Value Equation . . . 52
- 3.2 Fokker-Planck Equation for Behavioral Predictions . . . 54
  - 3.2.1 Determining x_B . . . 55
  - 3.2.2 Time scale of the FPE . . . 55
- 3.3 Evaluation of $\log[h(g)]$ . . . 55

4 Sequential dependencies across trials **56**
- 4.1 Biased passive policy . . . 57
- 4.2 Biased passive policy with biased rewards . . . 57
- 4.3 Biased prior beliefs . . . 58

5 KL control and exploration **58**

Here we present the mathematical derivation of our framework. To make the document self-contained, we include some theoretical background when needed.

1 Framework

1.1 Markov Decision Processes: First exit formulation

We consider a Markov Decision Process[19,35,44] (MDP) with a set of states $s \in S$ and a set of admissible actions $a \in A$. The actions generate transitions between states according to a transition probability $T(s'|s,a)$ which crucially is Markov in the states and actions. Associated to each transition, there is an immediate reward $\mathcal{R}(s,a)$. The goal of the agent is to maximize the long term reward accumulated over the entire sequence of transitions. In order to do this, the agent is equipped with a control policy $u(a|s)$, i.e., a decision rule to select actions in each state. The problem is then to find the optimal policy that ensures the largest possible long term reward.

In the "first-exit" formulation of the problem[35,44], there is a set $S_T \in S$ of "terminal" states which, once reached, terminate the process. The accumulated reward starting from state s and acting optimally thereafter, called the value function $Y(s)$, can be written as

$$Y(s) = \max_u E\left[\sum_{t=1}^{t_f} \mathcal{R}_t(s)\right] \tag{5}$$

where $\mathcal{R}_t(s)$ is the reward obtained at time t having started from state s and following the policy, and the expectation is taken over any stochasticity in the process. t_f denotes the time step at which the first terminal state is reached (which is also a random variable). The value function follows a recursive relationship called the Bellman Equation[35,107] (BE), which is given by

$$\begin{aligned} Y(s) &= \max_u \left\{ \sum_a u(a|s) \sum_{s'} T(s'|s,a)[\mathcal{R}(s,a) + Y(s')] \right\} \\ &= \max_u \left\{ \sum_a u(a|s) \left[\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) Y(s') \right] \right\} \end{aligned} \tag{6}$$

It can be shown that if the immediate cost $\mathcal{R}(s,a)$ does not depend on u, the optimal policy must be deterministic[19] and given by

$$u(a|s) = \arg\max_a \left\{ \mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) Y(s') \right\} \tag{7}$$

which allows rewriting the equation (6) as

$$Y(s) = \max_a \left\{ \mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) Y(s') \right\} \tag{8}$$

1.2 Kullback-Leibler control

In the above formulation of the problem, standard in the field of Reinforcement Learning, there are no costs associated to the control of the agent. In the field of Optimal Control, two reasonable features are added: First, it is assumed that the agent is capable of behaving in the absence of

control. Second, it is assumed that control is costly, and that this cost should be included in the optimization process. Formalizing this idea means considering *two* policies. One is a control-free default (or passive) policy $p(a|s)$, which describes the behavior of the agent in the absence of control and which, in principle, bears no particular relationship to the goals of the agent. This passive policy is part of the specification of the problem and plays a role similar to that of the prior in statistical inference. The other one is the target optimal policy $u(a|s)$ which describes the trade-off between long-term reward maximization and control costs.

The framework of KL control[34–36,40,42] assumes a specific form for the cost of control, which is added to the immediate reward $\mathcal{R}(s,a)$. The immediate consequence of choosing action a in state s under the optimal policy $u(a|s)$ for an agent with default policy $p(a|s)$ thus becomes

$$\mathcal{L}(s,a;p,u) = \mathcal{R}(s,a) - \beta^{-1}\log\left(\frac{u(a|s)}{p(a|s)}\right) \tag{9}$$

The constant β^{-1} measures the magnitude of the cost of control relative to the immediate reward, and can be considered a property of the agent. As we will see, when the cost of control is negligible (i.e., $\beta \to \infty$), the attainment of long term reward dominates the behavior of the agent, and the optimal policy becomes identical to the one obtained in the standard MDP framework. However, when control is costly, the optimal behavior of the agent will tend to be constrained by its passive policy. Replacing $\mathcal{R}(s,a)$ by $\mathcal{L}(s,a;p,u)$ in the BE (8) one obtains

$$\begin{aligned}
Y(s) &= \max_u \left\{ \sum_a u(a|s)\left[\mathcal{L}(s,a;p,u) + \sum_{s'} T(s'|s,a)Y(s')\right]\right\} \\
&= \max_u \left\{ \sum_a u(a|s)\left[\mathcal{R}(s,a) - \beta^{-1}\log\left(\frac{u(a|s)}{p(a|s)}\right) + \sum_{s'} T(s'|s,a)Y(s')\right]\right\} \\
&= \max_u \left\{ \sum_a u(a|s)\left[\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a)Y(s')\right] - \beta^{-1}KL\left(u(a|s)\,||\,p(a|s)\right)\right\}
\end{aligned} \tag{10}$$

where $KL(p||q)$ is the Kullback-Leibler divergence[41] between distributions $p(x)$ and $q(x)$, which gives its name to this framework. Thus, the effective consequence of the control cost is to penalize the optimal policy by a quantity proportional to the dissimilarity between the two policies, as reasonably expected. The usefulness of this formulation can be appreciated when applying some further algebra on the above equation:

$$\begin{aligned}
Y(s) &= \max_u \left\{ -\beta^{-1}\sum_a u(a|s)\log\left[\frac{u(a|s)}{p(a|s)\exp\left(\beta\mathcal{R}(s,a) + \beta\sum_{s'} T(s'|s,a)Y(s')\right)}\right]\right\} \\
&= \beta^{-1}\min_u \left\{ KL\left(u(a|s)\,\middle|\middle|\,\frac{p(a|s)\exp\left(\beta\mathcal{R}(s,a) + \beta\sum_{s'} T(s'|s,a)Y(s')\right)}{Z(s)}\right)\right\} + \beta^{-1}\log\left(Z(s)\right)
\end{aligned}$$

where the "partition function" $Z(s)$ is defined as

$$Z(s) = \sum_a p(a|s)\exp\left(\beta\left(\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a)Y(s')\right)\right) \equiv \sum_a u_u(a|s) \tag{11}$$

The control law results trivially from the minimization of the KL divergence, which is attained when its two arguments are equal, resulting in $KL = 0$. In this case, the BE reduces to

$$Y(s) = \beta^{-1} \log \left(Z(s) \right) = \beta^{-1} \log \left[\sum_a p(a|s) \exp \left(\beta \left(\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) Y(s') \right) \right) \right] \tag{12}$$

which is a self-consistency equation that specifies the value function. At the same time, the optimal policy is given by (from the condition $KL = 0$)

$$u(a|s) = \frac{u_u(a|s)}{\sum_a u_u(a|s)} = p(a|s) \frac{\exp \left[\beta \left(\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) Y(s') \right) \right]}{Z(s)} \tag{13}$$

The last two equations (the equivalent of equations (8) and (7) above for the standard MDP) provide the solution to the KL control problem. It entails two important simplifications: First, unlike Eq. (8), Eq. (12) does not have a $\max_a$ operator and it is thus continuous. Second, once the value function in Eq. (12) is known, Eq. (13) provides an analytical expression for the optimal policy. This expression, furthermore, has an intuitive interpretation. Except for the normalization constant $Z(s)$, the (un-normalized) optimal probability of choosing each action under the optimal policy $u_u(a|s)$ is proportional to the same probability under the passive policy $p(a|s)$, with a proportionality constant that grows exponentially the net "action value" of the pair (a, s). One important consequence of this fact is that actions that have zero probability of being chosen under the default policy remain forbidden for the optimal agent. Thus, in KL control, control can only bias the probability of actions that were possible by default, it cannot create new actions *de novo*. This limitation can be seen as a price to pay for the mathematical simplification associated to the choice of the KL divergence as a measure of control cost.

1.3 Equivalence between MDP and KL formulations

Equations (12) and (13) are instances of the functions LogSumExp and SoftMax respectively

$$\begin{aligned} \mathrm{LogSumExp}(\vec{x}) &= \log \left(\sum_i \exp(x_i) \right) \equiv \mathrm{LSe}(\vec{x}) \\ \mathrm{SoftMax}_i(\vec{x}) &= \frac{\exp(x_i)}{\sum_i \exp(x_i)} = \exp \left(x_i - \mathrm{LSe}(\vec{x}) \right) \equiv \vec{\mathrm{SM}}(\vec{x}) \end{aligned}$$

These are smooth versions of the *max* and *argmax* operators correspondingly. Notice that knowing the value of the first one facilitates the calculation of the second one, as it is the logarithm of the denominator. Expanding $Z(s)$ in both (13) and (12)

$$
\begin{aligned}
u(a|s) &= p(a|s)\frac{\exp\left[\beta\left(\mathcal{R}(s,a)+\sum_{s'}T(s'|s,a)Y(s')\right)\right]}{\sum_a p(a|s)\exp\left(\beta\left(\mathcal{R}(s,a)+\sum_{s'}T(s'|s,a)Y(s')\right)\right)} \\
&= \mathrm{SM}_a\left(\beta\left[\mathcal{R}(s,a)+\frac{1}{\beta}\log\left(p(a|s)\right)+\sum_{s'}T(s'|s,a)Y(s')\right]\right) \\
&= \mathrm{SM}_a(\beta\vec{Q}(s)) \qquad (14)
\end{aligned}
$$

$$
\begin{aligned}
Y(s) &= \beta^{-1}\log\left(\sum_a p(a|s)\exp\left(\beta\left[\mathcal{R}(s,a)+\sum_{s'}T(s'|s,a)Y(s')\right]\right)\right) \\
&= \beta^{-1}\mathrm{LSe}\left(\beta\left[\mathcal{R}(s,a)+\frac{1}{\beta}\log\left(p(a|s)\right)+\sum_{s'}T(s'|s,a)Y(s')\right]\right) \\
&= \beta^{-1}\mathrm{LSe}(\beta\vec{Q}(s)) \qquad (15)
\end{aligned}
$$

where we have defined the vector $\vec{Q}(s)$ with components

$$
Q_a(s) \equiv \mathcal{R}(s,a)+\frac{1}{\beta}\log\left(p(a|s)\right)+\sum_{s'}T(s'|s,a)Y(s') \qquad (16)
$$

Each $Q_a(s)$ measures the *action value* of being in state s and performing action a including a measure of control given by the surprisal $\log(p(a|s))$. Since this is an immediate cost, it is useful to include it in the immediate reward, which we thus redefine to be

$$
\tilde{\mathcal{R}}(s,a) = \mathcal{R}(s,a)+\frac{1}{\beta}\log\left(p(a|s)\right) \qquad (17)
$$

The second term is always negative, i.e., a cost. Thus, the more unlikely an action is in a particular state under the default policy, the lower the effective immediate reward in that state. This can be construed as a form of directed exploration towards the default. Using this modified effective immediate reward leads to the following natural expression for net action value $Q_a(s)$

$$
Q_a(s) \equiv \tilde{\mathcal{R}}(s,a)+\sum_{s'}T(s'|s,a)Y(s')
$$

It's straightforward to show that in the limit where β goes to infinity, $\vec{\mathrm{SM}}(\vec{x}) \to \arg\max(\vec{x})$, $\mathrm{LSe}(\vec{x}) \to \max(\vec{x})$, and $\tilde{\mathcal{R}}(s,a) \to \mathcal{R}(s,a)$, so that the value function and the policy converge to the MDP deterministic solution in Eqs. (7-8).

$$
\begin{aligned}
\lim_{\beta\to\infty} u(a|s) &= \arg\max_a\left\{\mathcal{R}(s,a)+\sum_{s'}T(s'|s,a)Y(s')\right\} \\
\lim_{\beta\to\infty} Y(s) &= \max_a\left\{\mathcal{R}(s,a)+\sum_{s'}T(s'|s,a)Y(s')\right\}
\end{aligned}
$$

1.4 Average-adjusted reward formulation

The first-exit formulation describes well a single trial in a decision making experiment, starting with stimulus onset and ending when the subject commits. However, a behavioral session contains many such trials, and requires a longer horizon. The standard way to deal with long (infinite) horizons is to use discounted rewards[19]. However, if the MDP is ergodic, there exists another approach which is well suited for describing a multi-trial decision-making experiment: reward rate maximization[44,45]. We explain this framework for a standard MDP, and later on show that it is trivially generalized to the KL control framework.

Let the MDP be represented by a discrete time, ergodic Markov chain[44]. We focus on the case where trajectories along the chain can be broken down in "trials", with the beginning of each trial marked by the moment in which a certain initial state S_I is visited. S_I should be understood as a reference state, and the ergodicity of the chain allows us to make this choice arbitrarily.

Let us define

$$Y_N(s) = E\left[\sum_{t=0}^{N-1} \mathcal{R}_t(s_0) \,\middle|\, s_0 = s\right]$$

with the same interpretation as in (5) except that N is a fixed and sufficiently large natural number, and where we have omitted the explicit dependence of the sequence of rewards on the policy (we in general assume the optimal policy). Since the state S_I is always revisited, we can rewrite the equation separating it in two terms: one that captures the accumulated reward starting from state s until S_I is reached for the first time, and another one capturing the remainder of the accumulated value since the first time S_I is visited

$$\begin{aligned} Y_N(s) &= \tilde{Y}(s) + Y_M \\ \tilde{Y}(s) &\equiv E\left[\sum_{t=0}^{k_1} \mathcal{R}_t(s_0) \,\middle|\, s_0 = s\right] \\ Y_M &= E\left[\sum_{l=1}^{M-1} \sum_{t=k_l}^{k_{l+1}} \mathcal{R}_t(s_{k_l}) \,\middle|\, s_{k_l} = S_I\right] \end{aligned} \tag{18}$$

where k_l corresponds to the successive times that the state S_I is revisited, and the index l goes over trials, which go up to M. The conditioning on $s_{k_l} = S_I$ makes explicit that every trial is initiated at the state S_I. Only the first term depends on the particular initial state s at time $t = 0$, while the second term is independent of s and linear in M. This is equivalent to separating the initial (in)finite horizon problem into an initial first-exit problem, and a reduced (in)finite horizon problem that consists of a collection of identical first exit problems.

The reward rate ρ is defined as

$$\rho = \lim_{N\to\infty} \frac{1}{N} E\left[\sum_{t=0}^{N-1} \mathcal{R}_t(s_0) \,\middle|\, s_0 = S_I\right] \tag{19}$$

Since the chain is ergodic, the reward rate for any policy is constant and independent of the initial state, so we can choose $s = S_I$. We can employ the same sort of separation of N into trials that we did above. Multiplying and diving by M in both numerator and denominator we obtain

$$\rho = \lim_{M\to\infty} \frac{M \cdot E\left\{ \frac{1}{M}\sum_{l=0}^{M-1}\sum_{t=k_l}^{k_{l+1}} \mathcal{R}_t(s_{k_l}) \,\middle|\, s_{k_l} = S_I \right\}}{M \cdot E\left\{ \frac{1}{M}\sum_{l=0}^{M-1} (k_{l+1} - k_l) \right\}} \tag{20}$$

The expectation in the numerator corresponds to the average cumulative reward in one trial, while the expectation in the denominator corresponds to the average duration of a trial. Let us call these $\langle \mathcal{R} \rangle$ and $\langle T \rangle$ respectively. Then we can simply write

$$\rho = \frac{\langle \mathcal{R} \rangle}{\langle T \rangle} \tag{21}$$

The reward rate ρ is also defined as the gain of the policy, since it measures the slope of the asymptotic linear dependence of total reward with the duration of the trajectory. Policies that obtain differences in total future reward which are constant as N grows will all have the same ρ, but it would be desirable to know which policies maximize these finite contributions as well. The relative value, or bias, $V(s)$ of the policy, defined as

$$V(s) \equiv \lim_{N\to\infty} E\left[\sum_{t=0}^{N-1} (\mathcal{R}_t(s_0) - \rho) \,\middle|\, s_0 = s \right] \tag{22}$$

measures exactly this finite contribution, and is thus dependent on the initial state s. It can be shown[44,45] that a policy that maximizes the relative value $V(s)$ also automatically maximizes the reward rate ρ. In a multi-trial decision problem, $V(s)$ can be written as

$$\begin{aligned} V(s) &= E\left[\sum_{t=0}^{k_1} (\mathcal{R}_t(s_0) - \rho) \,\middle|\, s_0 = s \right] + \lim_{M\to\infty} E\left[\sum_{l=1}^{M-1} \sum_{t=k_l}^{k_{l+1}} (\mathcal{R}_t(s_{k_l}) - \rho) \,\middle|\, s_{k_l} = S_I \right] \\ &= \tilde{Y}(s) - E\Big[k_1 | s_0 = s\Big]\rho + \lim_{M\to\infty} (M-1) \left(E\left[\frac{1}{M-1} \sum_{l=1}^{M-1} \sum_{t=k_l}^{k_{l+1}} \mathcal{R}_t(s_{k_l}) \,\middle|\, s_{k_l} = S_I \right] - \rho \langle T \rangle \right) \end{aligned} \tag{23}$$

The expectation in the third term corresponds to $\langle \mathcal{R} \rangle$ as we saw above, so that

$$V(s) = \tilde{Y}(s) - E\Big[k_1 | s_0 = s\Big]\rho + \lim_{M\to\infty} (M-1) \left(\langle \mathcal{R} \rangle - \rho \langle T \rangle \right) = \tilde{Y}(s) - E\Big[k_1 | s_0 = s\Big]\rho \tag{24}$$

where the last equality uses the definition of ρ in Eq. (21). We now write the BE for the value function, taking the limit of N to infinity

$$\lim_{N\to\infty} Y_N(s) = \max_u \left\{ \sum_a u(a|s) \left[\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) \lim_{N\to\infty} Y_N(s') \right] \right\}$$

which, recalling that $\lim_{N\to\infty} Y_N(s) \equiv \tilde{Y}(s) + Y_M$, and that Y_M is a constant independent of s, becomes

$$\tilde{Y}(s) = \max_u \left\{ \sum_a u(a|s) \left[\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) \tilde{Y}(s') \right] \right\} \tag{25}$$

Using Eq. (24), this can be turned into a BE for the relative value $V(s)$

$$V(s) + E\big[k_1|s_0 = s\big]\rho = \max_u \left\{ \sum_a u(a|s) \left[\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a)V(s') + \Big(E\big[k_1|s_0 = s\big] - 1\Big)\rho \right] \right\}$$

The last term uses the fact that the expected first exit time following a transition is exactly one unit less than the expected time before the transition, i.e.,

$$\sum_{s'} T(s'|s,a)E\big[k_1|s_0 = s'\big] = E\big[k_1|s_0 = s\big] - 1$$

Since the expectation in this equation is already over trajectories obtained using the optimal policy, it can be taken outside the max operator, which leads to

$$V(s) + \rho = \max_u \left\{ \sum_a u(a|s) \left[\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a)V(s') \right] \right\} \tag{26}$$

We have thus transformed an infinite horizon problem on Y into a first exit problem on V, with the additional constant ρ in the left hand side. Although $V(s)$ is a relative value, we will refer to it as the value function from now on. The value of ρ is unknown, but we know it has to satisfy Equation (24). This implies that $V(S_I)$ needs to satisfy the following condition

$$V(S_I) = \tilde{Y}(S_I) - E\Big[k_1|s_0 = S_I\Big]\rho$$

since S_I marks the start of a trial, the first term is the expected reward during one trial, $\langle \mathcal{R} \rangle$, and the expectation in the second term is equivalent to the average duration of one trial, $\langle T \rangle$. Thus

$$V(S_I) = \langle \mathcal{R} \rangle - \langle T \rangle \rho = 0 \tag{27}$$

In practice, this equation is used to obtain the value of ρ self-consistently: We solve the BE (26) for a fixed value of ρ, use it to evaluate $V(S_I)$, and repeat this procedure adjusting ρ iteratively until $V(S_I) = 0$. The value of ρ for which this condition is met is the average reward rate of the policy.

Notice that, mathematically, the BE (26) is identical to the BE for a first-exit problem with a "cost of time" ρ, i.e., a problem with an immediate reward for each (s,a) equal to $\mathcal{R}(s,a) - \rho$. Such BE is well defined for any arbitrary ρ. It is only if one wants to interpret $V(s)$ as the relative value of an infinite horizon problem with asymptotic reward rate ρ that Eq. (27) needs to be satisfied.

Finally, because no assumptions have been made on the form of the immediate reward, the average-adjusted reward formulation for KL control is trivially obtained replacing $\mathcal{R}(s,a)$ in the previous equations by Equation (9). The relevant expressions become

$$\begin{aligned}
u(a|s) &= \frac{\exp\left(\beta Q_a\right)}{\sum_a \exp\left(\beta Q_a\right)} = \mathrm{SM}_a(\beta \vec{Q}(s)) \\
V(s) &= \beta^{-1} \log\Big(\sum_a \exp\left(\beta Q_a\right)\Big) - \rho = \beta^{-1}\mathrm{LSe}(\beta \vec{Q}(s)) - \rho \\
Q_a(s) &\equiv \left[\mathcal{R}(s,a) + \frac{1}{\beta}\log\big(p(a|s)\big)\right] + \sum_{s'} T(s'|s,a)V(s')
\end{aligned}$$

1.5 Partial Observability: POMDPs

So far, we have been assuming that the agent has perfect knowledge about the states of the environment. In perceptual decision-making tasks, however, the challenge faced by the agent is precisely that the relevant states of the environment from the point of view of reinforcement are not directly observable, and have to be inferred through inference, based on stochastic observations. The appropriate mathematical framework to describe these situations is that of partially-observable Markov decision processes[20,108] (POMDPs).

Qualitatively, this introduces the need for "information seeking" actions. In the problem we describe in the text, this action corresponds to the postponement of commitment, or waiting (which we have denoted by W). Although accumulation of evidence across time would seem to require memory and thus violation of the Markov assumption, it can be shown that the probabilistic belief of the agent (i.e., the posterior probability of the states given the full history of observations) can be updated recursively in a Markovian fashion, i.e., the agent's belief in the current time-step is only a function of the current observation and the belief in the previous time-step[20,109]. Thus, formally, a POMDP can be construed as an MDP where states are replaced by beliefs[20]. If one uses the notation $\mathbf{b} = b(s_1), \ldots, b(s_{N_s})$ to refer to the relevant beliefs of the agent over the states s, then the POMDP formulation of our problem is very similar to the one we have previously described

$$u(a|\mathbf{b}) = \frac{\exp(\beta Q_a)}{\sum_a \exp(\beta Q_a)} = \mathrm{SM}_a(\beta \vec{Q}(\mathbf{b})) \tag{28a}$$

$$V(\mathbf{b}) = \beta^{-1} \log\left(\sum_a \exp(\beta Q_a)\right) - \rho = \beta^{-1}\mathrm{LSe}_a(\beta \vec{Q}(\mathbf{b})) - \rho \tag{28b}$$

$$Q_a(\mathbf{b}) \equiv \left[\sum_s b(s)\mathcal{R}(s,a) + \frac{1}{\beta}\log\left(p(a|\mathbf{b})\right)\right] + \sum_{\mathbf{b}'} T(\mathbf{b}'|\mathbf{b},a)V(\mathbf{b}') \tag{28c}$$

In this case, the immediate rewards $\mathcal{R}(s,a)$ now correspond to the expected reward by the agent given its current beliefs. The quantity $T(\mathbf{b}'|\mathbf{b},a)$ replaces the transition probability $T(s'|s,a)$ and describes the dynamics of belief induced by the dynamics of the agent-environment. Writing it in this form requires a marginalization over the observations that might be received if the agent performs action a with beliefs $\mathbf{b}$, which will determine its subsequent belief state $\mathbf{b}'$, as we now describe.

2 Modeling a binary choice over a continuous latent state

2.1 States and transitions

The states in the task follow a continuous time Markov Chain (CTMC) as depicted in Fig. 1. The most important state is the "stimulus" state. This state is the only one that admits actions from the agent, and therefore the only one that offers immediate rewards $\mathcal{R}(s,a)$. The possible actions are: waiting (W), choosing right (R) or choosing left (L).

The stimulus is characterized by a latent, continuous, unobserved feature (μ), which needs to be categorized by the agent as positive or negative. We define the task contingency to require R (L) when $\mu > 0$ ($\mu < 0$). We thus refer to $\mu > 0$ ($\mu < 0$) as the "right stimulus" – RS ("left stimulus"

– LS). If the agent waits, the current state is maintained (including the value of μ). But if the agent chooses R or L, the stimulus will end, and the task moves to a different state. In this transition, the agent will receive the a $\mathcal{R}(s, a)$ which depends on whether the choice is correct or an error. Let us define the reward for being correct as R_w and the reward for being incorrect as 0 (later we will show that the framework is invariant under the scaling of this reward). In addition, the following state will be different depending on a correct or an incorrect response. After a correct choice, the task moves directly to the "Inter Trial Interval" (ITI), with duration t_i. After an error, the task advances to a "Time Penalty" state (TP), with duration t_p, after which the task proceeds to the ITI. The time penalty becomes another incentive for the agent to choose correctly – besides the potential reward – as the goal is to maximize the reward rate.

We can construct a table to make the different state transitions and payoffs from the Stimulus state more explicit.

State transitions ($T(s'|s, a)$):

	R	L	W
RS	ITI	TP	RS
LS	TP	ITI	LS

Payoff matrix ($\mathcal{R}(s, a)$):

	R	L	W
RS	R_w	0	0
LS	0	R_w	0

Notice that the model of transition probabilities defined in this way is *deterministic*, as each combination of states and actions always gives raise to the same successor state.

2.2 Default policy

We model the agent's default response tendencies through a constant, stimulus-independent probability of commitment per unit time

$$p(R|s = \text{RS, LS}) = \phi\ \lambda dt \tag{29a}$$

$$p(L|s = \text{RS, LS}) = (1 - \phi)\lambda dt \tag{29b}$$

$$p(W|s = \text{RS, LS}) = 1 - \lambda dt \tag{29c}$$

where λ is the rate of commitment in Hz. In most cases we assume responding is unbiased, i.e., $\phi = 1/2$, although we relax this assumption when modeling sequential dependencies.

2.3 Inference

In this section we derive expressions for the temporal evolution of the agent's belief during the stimulus presentation. The relevant belief, g, is a scalar between 0 and 1 representing the probability that the stimulus is rightward, i.e., that $\mu > 0$. Since the latent state μ is continuous, the agent calculates g indirectly by first estimating the posterior distribution of μ given the observations. If at time t the posterior over μ is $p(\mu, t)$ then

$$g(t) = p(\text{RS}) = \int_0^\infty p(\mu, t) d\mu \tag{30}$$

The stimulus emits a stream of temporally uncorrelated Gaussian observations. It is simpler to initially assume an observation is emitted every short interval Δt with probability

$$p(o|\mu) = N(o|\mu\Delta t, \sigma^2\Delta t)$$

Without loss of generality, we also assume the prior distribution over μ is also Gaussian, and equal to

$$p_0(\mu) = N(\mu|0, \sigma_\mu) \tag{31}$$

which implies an initially neutral belief

$$g_0 = \int_0^\infty p_0(\mu)d\mu = \frac{1}{2} \tag{32}$$

Because both the prior and the likelihood are Gaussian, the posterior is always Gaussian. It is easy to show that the posterior after n observations is equal to[15]

$$p_n(\mu) \equiv p(\mu|o^{(1:n)}) = N\left(\mu\middle|\frac{x(t)}{t+t_g}, \frac{\sigma^2}{t+t_g}\right) \equiv p(\mu, t) \tag{33}$$

where

$$t \equiv n\Delta t \tag{34}$$

$$x^{(n)} \equiv \sum_{i=1}^{n} o^{(i)} \equiv x(t) \tag{35}$$

$$t_g \equiv \frac{\sigma^2}{\sigma_\mu^2} \tag{36}$$

and thus that the belief at this time is

$$g(t) \equiv \int_0^\infty p(\mu, t)d\mu = \Phi\left(\frac{x(t)}{\sigma\sqrt{t+t_g}}\right) \tag{37}$$

where Φ is the standard cumulative Gaussian distribution. These expressions show that the total accumulated evidence and elapsed time at a given moment are sufficient statistics to specify the posterior over the latent state, and thus also the agent's belief. Note also that when $t = t_g$ the posterior uncertainty is $\sigma_\mu^2/2$, which is half of its initial value. The constant t_g thus measures the characteristic timescale of the inference process.

It is convenient to renormalize the variables of the problem to make them dimensionless as follows

$$\mu \leftarrow \frac{\mu}{\sigma_\mu} \tag{38a}$$

$$t \leftarrow \frac{t}{t_g} \tag{38b}$$

$$x \leftarrow \frac{\sigma_\mu}{\sigma^2}x = \frac{x}{\sigma\sqrt{t_g}} \tag{38c}$$

which simplifies the expressions of the prior, posterior and belief to

$$p_0(\mu) = N(\mu|0,1) \tag{39a}$$

$$p(\mu,t) = N\left(\mu\middle|\frac{x(t)}{t+1},\frac{1}{t+1}\right) \tag{39b}$$

$$g(t) = \Phi\left(\frac{x(t)}{\sqrt{t+1}}\right) \tag{39c}$$

We also consider the problem in continuous time, by making the time interval infinitesimal, i.e., $\Delta t \to \delta t$, in which case the accumulated evidence $x(t)$ becomes a continuous Markov process[110] (CMP) with Langevin equation

$$\frac{dx}{dt} = \mu + \eta(t) \tag{40}$$

where $\eta(t)$ is a white noise process. To proceed, we need to obtain the posterior predictive distribution of $x(t+\delta t)$[111], meaning x in the following infinitesimal step, given $x(t)$. If μ was known, this would be simply:

$$p(x'(t+\delta t)|x(t),\mu) = N\big(x'(t+\delta t)\big|x(t)+\mu\delta t,\delta t\big)$$

However, the agent does not know μ, only its probability distribution $p(\mu,t)$. Marginalizing over μ one obtains

$$\begin{aligned} p(x(t+\delta t)|x(t),t) &= \int d\mu\; N\big(x(t+\delta t)\,|x(t)+\mu\delta t,\delta t\,\big)N\left(\mu\,\middle|\,\frac{x(t)}{t+1},\frac{1}{t+1}\right) \\ &= N\left(x(t+\delta t)\middle|x(t)\left(1+\frac{\delta t}{t+1}\right),\delta t\left(1+\frac{\delta t}{t+1}\right)\right) \end{aligned} \tag{41}$$

Thus, the prediction of $x(t)$ follows another CMP with equation

$$\frac{dx}{dt} = \frac{x(t)}{t+1} + \eta(t) \tag{42}$$

To obtain the corresponding (predictive) dynamics of belief, we write the last expression as a stochastic differential equation

$$dx = \mu_x(x,t)dt + \sigma_x(x,t)\partial B_t \tag{43}$$

with

$$\mu_x(x,t) = \frac{x(t)}{t+1} \tag{44a}$$

$$\sigma_x(x,t) = 1 \tag{44b}$$

and invoke Ito's lemma[112], using the fact that $g(t)$ is a monotonic function of $x(t)$ (Eq. 39c)

$$dg = \mu_g dt + \sigma_g\,\partial B_t = \left(\frac{\partial g}{\partial t} + \mu_x(x,t)\frac{\partial g}{\partial x} + \frac{\sigma_x^2}{2}\frac{\partial^2 g}{\partial x^2}\right)dt + \sigma_x\frac{\partial g}{\partial x}\partial B_t \tag{45}$$

After some algebra, one obtains that the terms inside the parenthesis of the RHS of Eq. 45 cancel each other so $\mu_g = 0$. Thus, the predictive distribution of the belief increment is normal and given by

$$p(g(t+dt)|g(t),t) = N\big(0, \sigma_g^2(g(t),t)\big) \tag{46}$$

with

$$\sigma_g^2(g(t),t) = \sigma_x \frac{\partial g}{\partial x} = \frac{e^{-\left[\Phi^{-1}(g(t))\right]^2}}{2\pi(t+1)} \tag{47}$$

2.4 Dimensional analysis

As just shown, it is natural to measure elapsed time since stimulus onset in units of t_g. Under an appropriate renormalization (Eq. 38), inference then becomes parameter-free. The remaining free parameters of the problem are λ, β, t_i, t_p and R_w. Constants with units of time or rate should be measured in units of t_g

$$\lambda \leftarrow t_g \lambda \tag{48a}$$

$$t_i \leftarrow \frac{t_i}{t_g} \tag{48b}$$

$$t_p \leftarrow \frac{t_p}{t_g} \tag{48c}$$

$$1 \leftarrow t_g \tag{48d}$$

Thus, λ, for instance, represents the expected number of default choices in the time interval t_g necessary to roughly estimate the stimulus, on average across difficulties. The control limitation β^{-1}, has units of reward, and so has the value function (Eq. 10). It is natural, thus, to use

$$\beta \leftarrow R_w \beta \tag{49a}$$

$$V \leftarrow \frac{V}{R_w} \tag{49b}$$

$$1 \leftarrow R_w \tag{49c}$$

Conceptually, the fact that β^{-1} has units of reward means that control limitations are to be measured relative to the motivation of the agent (see text).

Although the reward rate ρ is not a free parameter, the preceding arguments imply that the mathematical framework is invariant with respect to changes in the magnitude of t_g and R_w if one also defines

$$\rho \leftarrow \frac{t_g}{R_w} \rho \tag{50}$$

From now on, we always work with dimensionless parameters.

2.5 Bellman Equation

In order to apply the average-adjusted reward rate framework, one needs to select one state as a reference, preferably one that is fully observable. In a sensory discrimination task, the natural state

is the onset of the stimulus, which then has by definition a value of zero (Section 1.4). By backward induction, one can then obtain the value of the ITI and TP states. Since in these states there are no payoffs and the actions do not trigger any transitions, the value lost in those states is equal to the reward rate times their duration

$$V(ITI) = -\rho t_i \tag{51a}$$
$$V(TP) = -\rho(t_i + t_p) \tag{51b}$$

The agent transitions into these observable states from the stimulus state after committing to one of the two possible responses, with probabilties

$$\begin{aligned} T(ITI|g,t,R) &= g \\ T(ITI|g,t,L) &= 1-g \\ T(TP|g,t,R) &= 1-g \\ T(TP|g,t,L) &= g \end{aligned}$$

If the agent does not commit, it will remain in the stimulus state and its belief will evolve from the expected stimulus observations as described in the last section

$$T(\mathbf{b}'|\mathbf{b},W) \Leftrightarrow p(g'(t+dt)|g(t),t)$$

One then needs to compute the infinitesimal evolution of the expected value of the stimulus state. Because the predictive belief distribution (Eq. 46) depends explicitly on both g and t, the value of the stimulus state is also a function of belief and time since stimulus onset. The required expression is

$$\begin{aligned} \sum_{\mathbf{b}'} T(\mathbf{b}'|\mathbf{b},W)V(\mathbf{b}') &= \int_{g'} p(g'(t+dt)|g,t)V(g',t+dt) \\ &= \int_{g'} p(g'(t+dt)|g,t)V(g+dg,t+dt) \\ &= \int_{g'} p(g'(t+dt)|g,t)[V(g,t)+dV] \end{aligned} \tag{53}$$

The value differential can be computed using Ito's lemma again:

$$dV = \mu_V dt + \sigma_V \partial B_t = \left(\frac{\partial V}{\partial t} + \mu_g(g,t)\frac{\partial V}{\partial g} + \frac{\sigma_g^2}{2}\frac{\partial^2 V}{\partial g^2}\right)dt + \sigma_g \frac{\partial V}{\partial g}\partial B_t \tag{54}$$

Since we are just looking for the expectation, we need to concern ourselves only with the term μ_V, and since $\mu_g = 0$, we obtain

$$\mu_V = \frac{\partial V}{\partial t} + \frac{\sigma_g^2}{2}\frac{\partial^2 V}{\partial g^2}$$

which implies

$$\begin{aligned}\int_{g'} p(g'(t+dt)|g,t)V(g',t+dt) &= V(g,t) + \langle dV \rangle = V(g,t) + \mu_V dt \\ &= V(g,t) + dt\Big(\frac{\partial V}{\partial t} + \frac{\sigma_g^2}{2}\frac{\partial^2 V}{\partial g^2}\Big) \end{aligned} \tag{55}$$

These expressions permit an evaluation of the actions values Q_a, which in fact always appear exponentiated in the Bellman Eq. From Eq. 28c we can write

$$\begin{aligned}\exp\left[\beta Q_a\right] &= \exp\left[\beta\Big(\sum_s b(s)\mathcal{R}(s,a) + \sum_{\mathbf{b}'} T(\mathbf{b}'|\mathbf{b},a)V(\mathbf{b}') + \frac{1}{\beta}\log\left(p(a|\mathbf{b})\right)\Big)\right] \\ &\equiv p(a|\mathbf{b})\exp\left[\beta\widetilde{Q}_a\right] \end{aligned} \tag{56}$$

where

$$\begin{aligned}\widetilde{Q}_R &= \sum_s b(s)\mathcal{R}(s,R) + \sum_{\mathbf{b}'} T(\mathbf{b}'|\mathbf{b},R)V(\mathbf{b}') = g + gV(ITI) + (1-g)V(TP) \\ &= gR_w - g\rho t_i - (1-g)\rho(t_p + t_i) = B + D(2g-1) \\ \widetilde{Q}_L &= \sum_s b(s)\mathcal{R}(s,L) + \sum_{\mathbf{b}'} T(\mathbf{b}'|\mathbf{b},L)V(\mathbf{b}') = 1 - g + (1-g)V(ITI) + gV(TP) \\ &= (1-g)R_w - (1-g)\rho t_i - g\rho(t_p + t_i) = B - D(2g-1) \\ \widetilde{Q}_W &= \sum_s b(s)\mathcal{R}(s,W) + \sum_{\mathbf{b}'} T(\mathbf{b}'|\mathbf{b},W)V(\mathbf{b}') = \int_{g'} p(g'(t+dt)|g,t)V(g',t+dt) \\ &= V(g,t) + dt\Big(\frac{\partial V}{\partial t} + \frac{\sigma_g^2}{2}\frac{\partial^2 V}{\partial g^2}\Big) \end{aligned} \tag{57}$$

and where we have defined

$$B \equiv \frac{(1-\rho t_i) + (-\rho[t_p + t_i])}{2} = \frac{1 - \rho(2t_i + t_p)}{2} \tag{58a}$$

$$D \equiv \frac{(1-\rho t_i) - (-\rho[t_p + t_i])}{2} = \frac{1 + \rho t_p}{2} \tag{58b}$$

The quantity B represents the average immediate value across the two choices, and D represents their "contrast". We can now write an expression for the Bellman Eq. (28b) in continuous time

$$\begin{aligned}V(g,t) + \rho dt &= \beta^{-1}\log\Big(\sum_a \exp\left(\beta Q_a\right)\Big) \\ &= \beta^{-1}\log\Big(p(W|s)\exp\left(\beta\widetilde{Q}_W\right) + p(R|s)\exp\left(\beta\widetilde{Q}_R\right) + p(L|s)\exp\left(\beta\widetilde{Q}_L\right)\Big) \\ &= \beta^{-1}\log\Big((1-\lambda dt)\exp\left(\beta\widetilde{Q}_W\right) + \lambda dt\exp\left(\beta B\right)h(g)\Big) \\ &= \widetilde{Q}_W + \beta^{-1}\log\Big(1 - \lambda dt\big[1 - h(g)\exp\left(\beta(B - \widetilde{Q}_W)\right)\big]\Big) \end{aligned} \tag{59}$$

where we have defined

$$h(g) \equiv \cosh\big(\beta D(2g-1)\big) \tag{60}$$

Substituting the expression for $\widetilde{Q}_W$ in Eq. 57, and developing to first order in dt we obtain

$$\begin{aligned} V(g,t) + \rho dt = V(g,t) \quad &+ \quad dt\Big(\frac{\partial V}{\partial t} + \frac{\sigma_g^2}{2}\frac{\partial^2 V}{\partial g^2}\Big) \\ &- \quad \frac{\lambda}{\beta} dt\big[1 - h(g)\exp\big(\beta(B - V(g,t))\big)\big] \end{aligned}$$

which leads to

$$0 = \frac{\partial v}{\partial t} + \frac{\sigma_g^2}{2}\frac{\partial^2 v}{\partial g^2} + \frac{\lambda}{\beta}\Big(h(g)\exp[-\beta v] - 1\Big) - \rho \tag{61}$$

This is a second order, non-linear partial differential equation which $v(g,t) \equiv V(g,t) - B$ (the value function during the stimulus presentation relative to the action-value B of a completely uncertain choice) must satisfy. To solve it, one needs to specify initial and boundary conditions.

2.5.1 Initial and boundary conditions

The "initial" condition is in fact the value of the function at an unknown, large time T, because Eq. (61) needs to be solved backwards in time. This is the standard approach in Dynamic Programming, as the value of a state depends on the value of the options available in the future, not in the past. A first approximation can be obtained by considering the limit of Eq. (61) as $t \to \infty$. In this limit, v should be stationary. It might seem counterintuitive that v, which represents the value of the stimulus state during which the agent is, by definition, uncommitted, can acquire a stationary value, given that the agent is paying a cost of time ρ. Intuitively, this is possible because after a sufficiently long time since stimulus onset, the probability that the agent will not have committed is vanishingly small. The value of being uncommitted at that state is thus given by the consequences of an immediate choice, which do not depend on the time elapsed since stimulus onset. Thus, one can impose $\partial_t v = 0$. Since $\sigma_g^2 \to 0$ as $t \to \infty$ (Eq. 47), the Bellman Eq. 61 in this limit becomes algebraic

$$\frac{\lambda}{\beta}\Big(h(g)\exp[-\beta v(g)] - 1\Big) - \rho = 0$$

The steady-state value, which we denote by $\tilde{v}(g)$ is thus given by

$$\tilde{v}(g) = \beta^{-1}\left(\log(h(g)) + \log\left(\frac{\lambda}{\lambda + \beta\rho}\right)\right) \tag{62}$$

Qualitatively, the first term in this Eq. represents the value associated to an immediate choice, and the second term, which is negative, represents the loss in value – driven by the control limitations of the agent – derived from not being able to instantaneously choose the most valuable option, a delay that is always costly as long as ρ is non-zero. As such, this control cost vanishes when $\lambda \to \infty$ (because then the agent spontaneously chooses without delay) or when $\beta \to \infty$ (because in this case there are no control limitations).

In principle, one can initialize the value function at a sufficiently long time T as $\tilde{v}(g)$ and numerically iterate backwards (see below). But this procedure is inefficient because the value of T

necessary for $v(g,T) \sim \tilde{v}(g)$ is very large. One can make this process more efficient by finding the first perturbative correction to $\tilde{v}(g)$. We seek a singular perturbative expansion of the form

$$v(g,t) = \tilde{v}(g) + \epsilon u(g,t')$$

where $t' \sim O(1)$ and $\epsilon \equiv t'/t$. The function $u(g,t')$ describes the asymptotic temporal evolution of the value function discounting for the change of scale necessary to focus on infinitely long times. Changing variables in the Bellman Eq. 61 from t to t' and neglecting terms $\sim O(\epsilon^2)$ one obtains, after some algebra

$$u(g,t') = \frac{1}{t'} \frac{\beta 4 D^2}{(\lambda + \beta\rho) 2\pi} \frac{e^{-\left[\Phi^{-1}(g)\right]^2}}{h^2(g)} \tag{63}$$

which leads to our desired result

$$v(g,T) \equiv \tilde{v}(g) + \epsilon u(g,T) = \tilde{v}(g) + \frac{1}{T} \frac{\beta 4 D^2}{(\lambda + \beta\rho) 2\pi} \frac{e^{-\left[\Phi^{-1}(g)\right]^2}}{h^2(g)} \tag{64}$$

One can use this equation to provide an estimate for the long time T at which this asymptotic solution is expected to be accurate, by imposing that the overall scale of the second term is sufficiently small compared to the scale of the first one. Because u is maximal at $g = 0.5$ and the range of $\tilde{v}$ corresponds to $\tilde{v}(g=1) - \tilde{v}(g=0.5)$, we impose that

$$\frac{\tilde{u}(g=0.5,T)}{\tilde{v}(g=1) - \tilde{v}(g=0.5)} = \frac{1}{T} \frac{\beta^2 D^2}{(\lambda + \beta\rho)\pi \log(\cosh(\beta D))} < \delta$$

where $\delta \ll 1$ is a numerical tolerance (we typically choose $\delta = X$). Thus, we find

$$T = \frac{\beta^2 D^2}{\delta \pi (\lambda + \beta\rho) \log(\cosh(\beta D))} \tag{65}$$

In practice, our numerical solution for the Bellman Eq. proceeds as follows. At first, given δ, we compute T using Eq. 65, and define $T' = 1.2\ T$. We then compute the asymptotic value function $v(g,T')$ using Eq. 64. In order for T and T' to be adequate, we impose two conditions: (*i*) the numerically obtained value function $v(g,t)$ for $0 < t < T$ should be the same when using $v(g,T')$ or $v(g,1.2\,T')$ as the asymptotic solution, and (*ii*) we update T to be equal to the numerically obtained (see below) 0.999 quantile of the reaction time distribution of the policy. These conditions guarantee that the asymptotic solution is truly asymptotic and does not affect any of the results, and that the numerical procedure is efficient in the sense of not calculating $v(g,t)$ for values of time that the policy never visits. Usually, satisfying these solutions requires modifying iteratively the values of T and T' a few times.

The boundaries correspond to the values $g = 0$ and $g = 1$, meaning full certainty that the stimulus state is SR or SL respectively. In these conditions, barring limitations in control, the agent should always choose immediately, as we argued in deriving $\tilde{v}(g)$. Since, as we showed in Eq. 62, the cost induced by these limitations is time independent, the value function in conditions of certainty is constant and equal to $\tilde{v}(g=0,1)$

$$v(g=0,t) = v(g=1,t) = \tilde{v}(g=0,1) = \beta^{-1}\left(\log(\cosh(\beta D)) + \log\left(\frac{\lambda}{\lambda + \beta\rho}\right)\right) \tag{66}$$

2.5.2 Calculation of reward rate

As explained above, the reward rate ρ must be calculated by imposing that the relationship $V_0 \equiv V(g_0, 0) = 0$ be satisfied. We do this by an iterative procedure, starting with an initial guess and changing ρ repeatedly using the bisection method until the initial condition for the value function is satisfied. We find that V_0 is monotonic in ρ and smooth, so the procedure is not problematic. Given that ρ has to satisfy

$$\rho = \frac{\langle Reward \rangle - \langle Cost \rangle}{\langle RT \rangle + (1 - Accuracy)t_p + t_i} \tag{67}$$

then, for the bisection method, we consider initial upper and lower bounds equal to

$$\rho_{\max} = \frac{1}{t_i} \qquad\qquad \rho_{\min} = \frac{1}{2\lambda + t_p + 2t_i} \tag{68}$$

The expression for $\rho_{\min}$ is the reward rate of an agent following the default policy, which is a lower bound for any policy with $\beta \neq 0$.

2.6 Optimal Policy

Once the value funcion $v(g, t)$ and ρ are known, it is possible to calculate the optimal policy from Eq. (28a)

$$u(a|\mathbf{b}) = \frac{\exp\left(\beta Q_a(\mathbf{b})\right)}{\sum_a \exp\left(\beta Q_a(\mathbf{b})\right)} = \mathrm{SM}_a(\beta \vec{Q}(\mathbf{b}))$$

The denominator in this expression is obtained from the definition of the value function

$$V(g, t) = \beta^{-1} \log\Big(\sum_a \exp\left(\beta Q_a(g, t)\right)\Big) - \rho dt$$

so that

$$\sum_a \exp\left(\beta Q_a(g, t)\right) = \exp(\beta(V(g, t) + \rho dt) \sim \exp(\beta V(g, t))[1 + \beta\rho dt] + \mathcal{O}(dt^2)$$

One can therefore write

$$u(a|g, t) = \frac{\exp\left(\beta(Q_a(g, t) - V(g, t))\right)}{1 + \beta\rho dt} \tag{69}$$

The quantity $A_a(g, t) = Q_a(g, t) - V(g, t)$ in the exponent is known as the *advantage*[19,113] of action a. It measures the excess or deficit in expected value from choosing an action relative to the average of all available actions following a given policy.

For the actions $a = R, L$, the exponential of the action-value is proportional to the instantaneous probability of commitment under the default policy (see, e.g., Eq. 59), and thus also proportional to dt. Using the expressions for $Q_{R,L}(g, t)$ in Eqs. 56-57, and developing to first order in dt one obtains

$$\begin{aligned} u\left(\begin{matrix} R \\ L \end{matrix}\middle| g, t\right) &= \frac{\lambda}{2} dt \exp\left(\beta(B \pm D(2g - 1) - V(g, t))\right) + \mathcal{O}(dt^2) \\ &= \frac{\lambda}{2} dt \exp\left(\beta(\pm D(2g - 1) - v(g, t))\right) \end{aligned} \tag{70}$$

which corresponds to Eq. 4 and Fig. 2 in the text. The policy for the remaining action W is now obtained by normalization, i.e., $u(W|g,t) = 1 - u(R|g,t) - u(L|g,t)$.

It's informative to write the expression for the optimal policy for each choice as the product of the probability of commitment (which we denote as $\bar{W}$), times the probability of choosing R, L given that a commitment has been made

$$u\left(\begin{matrix} R \\ L \end{matrix}\middle| g,t\right) = u(\bar{W}|g,t)u\left(\begin{matrix} R \\ L \end{matrix}\middle| \bar{W},g\right) \tag{71}$$

The probability of commitment is simply given by the sum of $u(R|g,t)$ and $u(L|g,t)$

$$u(\bar{W}|g,t) = \lambda\, dt\, h(g) \exp\left(-\beta v(g,t)\right) \tag{72}$$

which leads to

$$u\left(\begin{matrix} R \\ L \end{matrix}\middle| \bar{W},g\right) = \frac{1}{2}\Big(1 + \tanh\left(\beta(\pm D(2g-1))\right)\Big) \tag{73}$$

The complexity of the optimal policy is thus contained in the commitment decision, which depends on the time-dependent value of the stimulus state that requires solving the Bellman Eq. Once a commitment has been made, each action is selected through a simple, time-independent sigmoidal function of belief, which the steepness of the sigmoid an increasing function of the control ability β of the agent, and the "contrast" in value D between the two choices.

2.7 Accuracy, reaction time and decision confidence

The optimal policy depends on belief and time, but the subject's belief is unobservable. In each trial of an experiment, the subjects provide a binary choice, a reaction time and a measure of decision confidence. We seek a mathematical expression for the probability that, for a given stimulus difficulty μ, the subject will choose action $a = R, L$ at time $t = RT$ with a given level of confidence $g = DC$.

It turns out that it is simpler to work with the temporal evolution of accumulated evidence $x(t)$ than with the subject's instantaneous belief $g(t)$, which does not represent a problem since there is a one-to-one monotonic relationship between them (Eq. 39c). It is important to note that, for this analysis, the dynamics of the accumulated evidence $x(t)$ is given by Eq. 40 under a *known* stimulus μ (since μ is controlled by the experimenter), and not by the predictive distribution of $x(t)$ induced by Eq. 42. The predictive distribution is needed to compute the optimal policy (which only assumes prior knowledge of the stimulus distribution $p_0(\mu)$, Eq. 39a). Then, this policy is applied in all trials, under experiment-controlled sensory evidence characterized by μ.

We will slightly abuse notation and denote as $u_{R,L}(x(g),t)dt$ the optimal policy Eq. 70 as a function of x, and as $u_{\bar{W}}(x,t)dt$ the corresponding probability of commitment. Because $x(t)$ is a CMP, there is a Fokker-Plack Equation (FPE) associated with Eq. 40, which describes the dynamics of evidence accumulation in the absence of commitment

$$\frac{\partial}{\partial t}p(x,t;\mu) = -\mu\frac{\partial}{\partial x}p(x,t;\mu) + \frac{\partial^2}{\partial x^2}p(x,t;\mu)$$

In the presence of a choice policy, the distribution $p_{\bar{W}}(x,t;\mu)$, defined as the density at (x,t) given that the agent has not yet committed, will lose mass at a rate proportional to the rate of commitment

$u_{\bar{W}}(x,t)$ (Eq. 72), and will thus satisfy the following FPE

$$\frac{\partial}{\partial t}p_{\bar{W}}(x,t;\mu) = -\mu\frac{\partial}{\partial x}p_{\bar{W}}(x,t;\mu) + \frac{\partial^2}{\partial x^2}p_{\bar{W}}(x,t;\mu) - u_{\bar{W}}(x,t)p_{\bar{W}}(x,t;\mu) \tag{74}$$

The probability mass that escapes is precisely what is needed in order to compute the probability density of committing to action R, L at time t, with accumulated evidence x.

$$p_{R,L}(x,t;\mu) = u_{R,L}(x,t|\bar{W})u_{\bar{W}}(x,t)p_{\bar{W}}(x,t;\mu) = u_{R,L}(x,t)p_{\bar{W}}(x,t;\mu) \tag{75}$$

Eq. (74) needs to be solved under "natural" boundary conditions, since the domain is unbounded (see next section), and with initial condition

$$p(x,t=0) = \delta(0) \tag{76}$$

Given $p_{R,L}(x,t;\mu)$, all quantities of interest can be computed through marginalization. For instance, the psychometric function, RT distributions and chronometric functions are given by

$$P_{R,L}(\mu) = \int_0^\infty \int_{-\infty}^\infty p_{R,L}(x,t;\mu)\,dx\,dt \tag{77a}$$

$$p_{R,L}(t;\mu) = \frac{1}{P_{R,L}(\mu)}\int_{-\infty}^\infty p_{R,L}(x,t;\mu)\,dx \tag{77b}$$

$$\langle RT\rangle_{R,L}(\mu) = \int_0^\infty t\,p_{R,L}(t;\mu)\,dt \tag{77c}$$

Similar expressions can be obtained for the distributions of decision confidence and their means.

2.7.1 Decision lapses

We compute lapse rates by evaluating the value of $u(R,L|\bar{W},g)$ (Eq. 73) at $g=1$ and $g=0$ respectively. These are symmetrical for an unbiased default policy

$$\text{Lapse}_R = 1 - u\left(R|\,\bar{W},1\right) = 1 - \frac{1}{1+\exp\left(-2\beta D\right)} = \frac{1}{1+\exp\left(2\beta D\right)}$$
$$\text{Lapse}_L = u\left(R|\,\bar{W},0\right) = \frac{1}{1+\exp\left(2\beta D\right)}$$

Intuitively, although computing the psychometric function in general requires the procedure described in the last section, lapses correspond to the limit $\mu \to \pm\infty$, and in this limit the situation simplifies. The distribution of $g(t)$ effectively becomes a delta function at $g=0$ or $g=1$ instantaneously, and the rate is thus purely limited by the value of $u(R,L|\bar{W},g=0,1)$.

3 Numerical methods

3.1 Value Equation

We solve Eq. 61 using finite differences, in particular the Method of Lines (MoL)[114,115]. Here, one uses algebraic approximations to spatial derivatives, and solves the resulting algebraic system of ordinary differential equations (ODEs) using standard packages.

We approximate

$$\frac{\partial^2 v}{\partial^2 g} \sim \frac{v_{i+1} - 2v_i + v_{i-1}}{\Delta g^2}$$

where i is an index that designates a position along a grid in g and Δg is the spacing in g along the grid, which is assumed constant. We will consider that the grid as M points, so $i = 1$ and $i = M$ are the extremes. These points will correspond to the boundary conditions, and we will need to treat them separately. With this discretization, the value equation takes the following form

$$\frac{dv_i}{dt} = -\frac{\sigma_{g_i}^2}{2}\frac{v_{i+1} - 2v_i + v_{i-1}}{\Delta g^2} + \frac{\lambda}{\beta}\Big(1 - h_i \exp\left(-\beta v_i\right)\Big) + \rho\,, \quad 1 < i < M \tag{78}$$

with

$$\sigma_{g_i}^2 \equiv \sigma_g^2(g_i) \qquad\qquad h_i \equiv h(g_i)$$

For the boundary conditions we have

$$v_1 = v_M = \tilde{v}(g = 0) = \tilde{v}(g = 1) = \beta^{-1}\left(\log(\cosh(D)) + \log\left(\frac{\lambda}{\lambda + \beta\rho}\right)\right) \tag{79}$$

and the initual condition is just

$$\tilde{v}_i = \beta^{-1}\left(\log(h_i) + \log\left(\frac{\lambda}{\lambda + \beta\rho}\right)\right) \tag{80}$$

Eqs. 78 can be written in matrix form as

$$\frac{d\mathbf{v}}{dt} = \frac{r}{t+1}F \cdot A\mathbf{v} + \frac{\lambda}{\beta}\Big(1 - H\exp\left(-\beta\mathbf{v}\right)\Big) + \rho \tag{81a}$$

$$r = \frac{1}{4\pi\Delta g^2}\,, \quad F = Diag(\mathbf{f})\,, \quad H = Diag(\mathbf{h}) \tag{81b}$$

where

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$

and where we have defined

$$\sigma_g^2(g,t) = \frac{f(g)}{2\pi(t+1)} = \frac{e^{-\left[\Phi^{-1}(g(t))\right]^2}}{2\pi(t+1)}$$

We typically use $M = 501$.

3.2 Fokker-Planck Equation for Behavioral Predictions

Several techniques have been developed to maintain the positivity and the total probability of FPEs[116–118]. We use the Chang-Cooper method[116], providing only enough details to keep our description self-contained. We start by writing the FPE 74 in flux form. The presence of the sink term does not allow achieving the canonical flux form, but nonetheless we can write

$$\frac{\partial}{\partial t} f(x,t;\mu) = \frac{\partial}{\partial x} F(x,t;\mu) - u_{\bar{W}}(x,t) f(x,t;\mu)$$
$$F(x,t;\mu) = -\mu f(x,t;\mu) + \frac{1}{2}\frac{\partial}{\partial x} f(x,t;\mu)$$

Where we have changed the notation $f \equiv p_{\bar{W}}$, and F is the flux. Now we discretize x in a grid x_i such that $\Delta x = x_{i+1} - x_i$, with $x_{i\pm 1/2} \equiv x_i \pm \Delta x/2$, and consider the discretization of the flux derivative

$$\frac{\partial}{\partial x} F(x,t;\mu) \sim \frac{F_{i+1/2}(t) - F_{i-1/2}(t)}{\Delta x}$$

The goal is to choose the quantities $F_{i+1/2}(t)$ such that the positivity and normalization of the solution is guaranteed. This is achieved by setting

$$F_{i+1/2}(t) = \mu\Big[(1-\delta) f_{i+1}(t) + \delta f_i(t)\Big] + \frac{1}{2}\frac{f_{i+1}(t) - f_i(t)}{\Delta x}$$

where we have defined $f_i(t) \equiv f(x_i, t)$, and δ takes the value

$$\delta \equiv \frac{1}{2\Delta x\, \mu} + \frac{1}{1 - \exp(2\Delta x\, \mu)}$$

(see [117, 118] for details). This is valid for interior points. Regarding the boundary conditions, the physical domain of x is unbounded, but we have to limit it somehow in order to be able to compute the solution. The approach that we will use is to apply an absorbing barrier at $\pm x_B$, that is permeable to the sink force but captures the probability mass that diffuses into it. $x_B > 0$ is sufficiently large as to not disrupt the solution inside the barrier significantly[119], given that probability mass is escaping due to the sink term. This is implemented reducing the equation at the boundaries $i = 0$, $i = M$ to the following form

$$\frac{\partial}{\partial t} f_i(t) \sim -u_{\bar{W},i}(t) f_i(t)\ ,\ i = 0, M$$

The resulting algebraic equation in vector-matrix form is

$$\frac{d\mathbf{f}}{dt} = (A - U)\mathbf{f} \tag{82}$$

with A equal to

$$A = \begin{pmatrix} 0 & 0 & & & & & \\ 0 & a & c & & & & \\ & b & a & c & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & b & a & c & \\ & & & & b & a & 0 \\ & & & & & 0 & 0 \end{pmatrix} \qquad \begin{cases} a &= -\left(\dfrac{\mu(2\delta-1)}{\Delta x} + \dfrac{1}{\Delta x^2}\right) \\ b &= \dfrac{\mu\delta}{\Delta x} + \dfrac{1}{2\Delta x^2} \\ c &= \dfrac{\mu(\delta-1)}{\Delta x} + \dfrac{1}{2\Delta x^2} \end{cases}$$

and $U = Diag(\mathbf{u}_{\bar{\mathbf{W}}})$. As for the initial condition, it is somewhat problematic to numerically implement a Dirac delta. A practical approach is to approximate the sink term as constant for a brief time interval Δt, so then the PDE has an analytical solution:

$$f(x, \Delta t) = N(x|\mu\Delta t, \Delta t)\exp(-u_{\bar{W}}(0,0)\Delta t)$$

We typically use $\Delta x = 0.01$ and $\Delta t = 0.0002$.

3.2.1 Determining x_B

The limits x_B should be chosen such that the impact of using finite boundaries is negligible on the solution of the FPE. We choose x_B as to limit the maximum belief achieved. Given a certain belief cutoff defined as $1 - \delta g$, with $\delta g \ll 1$ (typically $\delta g = 10^{-3}$ will be reasonable), we have

$$x_B = \sqrt{T+1}\,\Phi^{-1}(1-\delta g) \tag{83}$$

Since it is more convenient to have a constant bound, we choose $t = T$ (Eq. 65) which provides us with the most conservative estimate.

3.2.2 Time scale of the FPE

We are already measuring time in units of t_g, such that it is adimensional. However, T still presents a large range of variation, so when solving the FPE it is convenient to measure time in units of T, such that it always goes from 0 to 1. In addition, if we define the x scale as $\sqrt{T}$, the FPE becomes[120]

$$\frac{\partial}{\partial t'} p_{\bar{W}}(x', t'; \mu) = -\mu\sqrt{\tau}\frac{\partial}{\partial x'} p_{\bar{W}}(x', t'; \mu) + \frac{1}{2}\frac{\partial^2}{\partial x'^2} p_{\bar{W}}(x', t'; \mu) - u_{\bar{W}}(x'\sqrt{T}, t'T)\tau p_{\bar{W}}(x', t'; \mu) \tag{84}$$

where $t' = t/T$, $x' = x/\sqrt{T}$. Therefore to convert to real time, we multiply t' by $t_g T$ (to undo both this latter change of variables and the initial one).

3.3 Evaluation of $\log[h(g)]$

The Bellman Eq. or the equations for the optimal policy contain terms of the form $h(g)\exp[-\beta v(g,t)]$, which we evaluate as

$$h(g)\exp[-\beta v(g,t)] = \exp[\log[h(g)] - \beta v(g,t)]$$

The quantity $\log[h(g)]$ can be numerically challenging. To avoid numerical instabilities, we write

$$\mathrm{LSE}(x,y) = \log\Big[\exp(x)+\exp(y)\Big] = \max(x,y) + \log\Big[1+\exp\big(\min(x,y)-\max(x,y)\big)\Big] \tag{85}$$

because the function $\mathrm{log1pexp} \equiv \log[1+\exp(x)]$ can be implemented achieving high numerical precision for $x \leq 0$. Then we can use the identities

$$\max(x,y) = \frac{1}{2}(x+y+|x-y|\,)$$
$$\min(x,y) = \frac{1}{2}(x+y-|x-y|\,)$$

to write

$$\mathrm{LSE}(x,y) = \frac{1}{2}(x+y+|x-y|\,) + \mathrm{log1pexp}\big(-|x-y|\,\big) \tag{86}$$

4 Sequential dependencies across trials

Although history effects generate long-timescale dependencies, where parameters in a particular trial might depend on events many trials in the past, their effect on the value function and optimal policy is tractable. This is, first, because in the average adjusted reward framework different trials only interact through the value of the reward rate, and second, because the effect of cross-trial dynamics on parameters only affects the reward rate through the steady state distribution of parameter values, as long as the dynamics of the parameters can be approximated as an ergodic Markov chain.

As described in the text, we consider sequential dependencies induced by across-trial dynamics in three parameters: a bias ϕ in the default policy (Eq. 29), the reward associated to correct responses to one of the options (which we will refer to as r), and the prior belief of the agent g_0 at the beginning of the trial. Globally, we refer to these are bias-inducing-parameters (BIPs). We define the vector $\mathbf{X}$ as containing all sets of BIPs that might be encountered in a session

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^T \qquad \text{with} \quad X_i = (\phi_i, r_i, g_{0_i})$$

Since the Bellman Equation for the whole session in the average adjusted reward framework is identical to the one for a single trial except for the specification and interpretation of the reward rate ρ (section 1.4), for any X_i we can calculate the associated value function $V_i(t,g;X_i)$, for a given ρ. One may interpret the X_i as extending the state space, which now includes the BIPs. The stimulus onset (i.e, $t=0$) still corresponds to a proper reference state (section 1.4). Thus, to determine the value of the reward rate ρ, we define a mixture of the $V_i(t,g;X_i)$ weighted by the stationary probability $\pi_S(X_i)$, and use the condition that the value V_0 of the mixture at $t=0$ should be equal to zero (section 1.4), i.e.

$$V_0 \equiv \sum_i \pi_S(X_i) V_i(t=0, g=g_{0_i}; X_i) = 0 \tag{87}$$

We find ρ using a similar iterative procedure as for the case without sequential dependencies. Since there is a single, constant ρ regardless of how many states n are part of $\mathbf{X}$, at each step all $V_i(X_i)$ are computed using the same tentative value of ρ to then re-evaluate Eq. (87) until convergence. Once the correct value of ρ has been found, then each V_i is re-evaluated, and we can directly compute

the policies for each X_i, which then allow deriving behavioral predictions (such as psychometric functions) for each BIP set.

In practice, to obtain the results in Figure 8, for each type of bias we consider BIP vectors of three states where only the correspondent parameter was changed in opposite directions, while the others were kept at their baselines:

$$\mathbf{X}_\phi = \Big((0.5, 1, 0.5), (0.5 + \Delta\phi^+, 1, 0.5), (0.5 + \Delta\phi^-, 1, 0.5)\Big)^T$$
$$\mathbf{X}_r = \Big((0.5, 1, 0.5), (0.5, 1 + \Delta r^+, 0.5), (0.5, 1 + \Delta r^-, 0.5)\Big)^T$$
$$\mathbf{X}_{g_0} = \Big((0.5, 1, 0.5), (0.5, 1, 0.5 + \Delta g_0^+), (0.5, 1, 0.5 + \Delta g_0^-)\Big)^T$$

The stationary probability distributions were in all cases $\pi_S = (1/3, 1/3, 1/3)^T$. The results do not change qualitatively for any reasonable choice of these probabilities. The procedure we just described is similar in spirit to a recently published method to categorize discrete behavioral states[103]. In fact, one can also include the parameters β and λ inside $\mathbf{X}$ and describe changes in control capability and impulsivity across trials. This should be addressed in future work.

4.1 Biased passive policy

The passive policy is now

$$p(W|s) = 1 - \lambda dt \tag{88a}$$
$$p(R|s) = \phi\lambda dt \tag{88b}$$
$$p(L|s) = (1 - \phi)\lambda dt \tag{88c}$$

such that ϕ represents the bias towards responding R. The magnitude of ϕ enters the value function through a modification of $h(g)$ in Eq. 60, which is now given by

$$h(g) = \phi \exp\big(\beta D(2g - 1)\big) + (1 - \phi) \exp\big(-\beta D(2g - 1)\big) \tag{89}$$

Given the new form for $h(g)$ value function, policies and behavioral predictions proceed as before.

4.2 Biased passive policy with biased rewards

We consider now a payoff matrix of the type

Payoff matrix ($\mathcal{R}'(s, a)$):

	R	L	W
RS	r	0	0
LS	0	1	0

This causes the action value $\widetilde{Q}_R$ to be rewritten as

$$\widetilde{Q}_R = \left(D + \frac{r-1}{2}\right)(2g - 1) + B + \frac{r-1}{2} \tag{90}$$

Therefore the function $h(g)$ now is

$$h(g) = \phi \exp\left[\beta\left(\frac{r-1}{2}\right)\right] \exp\left[\beta\left(D + \frac{r-1}{2}\right)(2g-1)\right] + (1-\phi)\exp\big(-\beta D(2g-1)\big) \tag{91}$$

from which again value function, policies and behavioral predictions proceed as before.

4.3 Biased prior beliefs

We implement the bias in the agent's prior belief about the stimulus as a shift in the distribution $p_0(\mu)$, which is effectively equivalent to a change in $g_0(t=0)$, following Equation 32. Due to the one-to-one mapping between g and x, a positive (negative) value of g_0 corresponds to a positive (negative) offset in the starting point of the evidence accumulation. Since the evolution of belief is unaffected by this change, none of the relevant equations need to be modified. One only needs to make sure that the value of the reference state used to calculate ρ is $V(t=0, g_0)$, and that the initial condition of the FPE (Eq. 76) is shifted accordingly, producing the observed bias in the psychometric functions.

5 KL control and exploration

As described in the text (see also sections 1.2 and 1.3), KL control modifies the standard sequential decision problem in two ways: Choices become exploratory, and this exploration is biased against actions that are unlikely given the default policy. In the standard KL framework, these two effects are coupled but, in general, they can be dissociated, as we now show. The problem began with a reformulation of the immediate reward of an action (Eq. 9)

$$\mathcal{L}(s,a;p,u) = \mathcal{R}(s,a) - \beta^{-1}\log\left(\frac{u(a|s)}{p(a|s)}\right) = \mathcal{R}(s,a) - \beta^{-1}\log(u(a|s)) + \beta^{-1}\log(p(a|s)) \tag{92}$$

This suggests a more general formulation where the two logarithmic terms are weighted differently

$$\mathcal{L}(s,a;p,u) = \mathcal{R}(s,a) - \beta_u^{-1}\log(u(a|s)) + \beta_p^{-1}\log(p(a|s)) \tag{93}$$

with β_u and β_p two independent parameters. Similar derivations as in sections 1.2 and 1.3 lead to the following solution for this more general case

$$u(a|s) = \frac{\exp(\beta_u Q_a)}{\sum_a \exp(\beta_u Q_a)} = \mathrm{SM}_a(\beta_u \vec{Q}(s)) \tag{94}$$

$$Y(s) = \beta_u^{-1}\log\Big(\sum_a \exp(\beta_u Q_a)\Big) = \beta_u^{-1}\mathrm{LSe}(\beta_u \vec{Q}(s)) \tag{95}$$

$$Q_a(s) \equiv \mathcal{R}(s,a) + \beta_p^{-1}\log\big(p(a|s)\big) + \sum_{s'} T(s'|s,a)Y(s') \tag{96}$$

The parameter β_u controls the steepness of the soft-max policy, whereas the parameter β_p controls the relative importance of the immediate reward $R(s,a)$ and the default policy in defining the goals of the agent. There are several interesting regimes for this problem

- When $\beta_p^{-1} \to 0$, the previous equations show that the optimal policy simply has a soft-max form parametrized by β_u. Unlike standard uses of soft-max exploration in reinforcement learning, however, in this scenario the value function and the (exploratory) policy are self-consistent. This formalism is known as Maximum Entropy Reinforcement Learning[42,121] (Max-Ent RL), because the agent is trying to maximize reward subject to an entropy constraint, as is evident from the form of the Bellman equation when $\beta_p^{-1} = 0$

$$Y(s) = \max_u \left\{ \sum_a u(a|s) \left[\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) Y(s') \right] + \beta^{-1} \mathcal{H}\left(u(a|s)\right) \right\}$$

  where

$$\mathcal{H}\left(u(a|s)\right) = -\sum_a u(a|s) \log[u(a|s)]$$

  Thus, "random" exploration can be construed as biased reward maximization, with a bias proportional to the surprisal of the optimal policy.

- In the other extreme, when $\beta_p \to 0$, the immediate reward of the agent effectively becomes independent of $\mathcal{R}(s,a)$ and equal to $\beta_p^{-1} \log(p(a|s))$. Here, the agent is 'locked' in the default policy, but its sensitivity to the likelihood of actions under the default depends on β_u. When β_u is large, the agent 'exploits' only actions which are likely under the default.

- In the general case, β_p measures the relative weight of the task-related rewards $\mathcal{R}(s,a)$ versus the default policy in shaping the agent's goals. It is easy to show that in this case, the 'unnormalized' optimal policy in Eq. 13 generalizes to

$$u_u(a|s) = p(a|s)^\alpha \exp\left[\beta_u \left(\mathcal{R}(s,a) + \sum_{s'} T(s'|s,a) Y(s')\right)\right] \tag{97}$$

  where $\alpha \equiv \beta_u / \beta_p$. This equation, together with Eq. 94 above, shows that the form of the task-independent, purely exploratory limit where $\beta_u \to 0$ and $\beta_p \to 0$ depends on which of these two parameters vanishes faster. When $\alpha = 1$, one recovers the KL control framework, and the limiting policy is the default

$$\lim_{\beta_u, \beta_p \to 0} u(a|s) = p(a|s)$$

  When $\alpha \sim 0$, the limiting policy becomes uniform

$$\lim_{\beta_u, \beta_p \to 0} u(a|s) = \frac{1}{|\mathcal{A}|}$$

  And when $\alpha^{-1} \sim 0$, the limiting policy is to always choose the most likely action under $p(a|s)$

$$\lim_{\beta_u, \beta_p \to 0} u(a|s) = \arg\max_a \left\{ p(a|s) \right\}$$

References

107. Bellman, R. Dynamic programming. *Princeton, USA: Princeton University Press* **1,** 3 (1957).

108. Monahan, G. E. State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms. *Management science* **28,** 1–16 (1982).

109. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).

110. Gillespie, D. T. *Markov processes: an introduction for physical scientists* (Elsevier, 1991).

111. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian data analysis* (Chapman and Hall/CRC, 1995).

112. Oksendal, B. *Stochastic differential equations: an introduction with applications* (Springer Science & Business Media, 2013).

113. Dayan, P. & Balleine, B. W. Reward, motivation, and reinforcement learning. *Neuron* **36,** 285–298 (2002).

114. Schiesser, W. E. *The numerical method of lines: integration of partial differential equations* (Elsevier, 2012).

115. Schiesser, W. E. & Griffiths, G. W. *A compendium of partial differential equation models: method of lines analysis with Matlab* (Cambridge University Press, 2009).

116. Chang, J. & Cooper, G. A practical difference scheme for Fokker-Planck equations. *Journal of Computational Physics* **6,** 1–16 (1970).

117. Pareschi, L. & Zanella, M. Structure preserving schemes for nonlinear Fokker–Planck equations and applications. *Journal of Scientific Computing* **74,** 1575–1600 (2018).

118. Mohammadi, M. Analysis of discretization schemes for Fokker-Planck equations and related optimality systems (2015).

119. Griffiths, G. & Schiesser, W. E. *Traveling wave analysis of partial differential equations: numerical and analytical methods with MATLAB and Maple* (Academic Press, 2010).

120. Langtangen, H. P. & Pedersen, G. K. *Scaling of differential equations* (Springer Nature, 2016).

121. Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., *et al. Maximum entropy inverse reinforcement learning.* in *Aaai* **8** (2008), 1433–1438.