# The metacognitive control of decisions predicts whether and how mice override their default policy

*K. Barc[1,2,*], C. Schreiweis[1,2,*], M. Euvrard[1], J. Mattout[2], E. Burguière[1,2,**], J. Daunizeau[1,3,**]*

[1] Paris Brain Institute - ICM, Hôpital de la Pitié Salpêtrière, Paris, France

[2] Lyon Neuroscience Research Center, Univ Claude Bernard Lyon 1, Inserm, CNRS, UMR 5292 / U-1028, Lyon, France

[3] ETH, Zurich, Switzerland

***,**** These authors contributed equallly to the work

Address for correspondence:

Jean Daunizeau

Motivation, Brain and Behavior Group

Paris Brain Institute (ICM)

47, bd de l'Hôpital, 75013, Paris, France.

Tel: +33 1 57 27 43 26

Mail: jean.daunizeau@icm-institute.org

Web: http://sites.google.com/site/jeandaunizeauswebsite

**ABSTRACT**

Decisions permeate every aspect of our lives (what to eat, where to live, etc) but decision-making policies seem to vary tremendously across time and across individuals. Rather than processing all decision-relevant information, we often rely on fast habitual and/or intuitive decision policies, which can lead to irrational biases and errors. Yet, we don't always follow the fast and negligent lead of habits or intuitions. So what determines whether we engage cognitive control and override our default responses? A possibility is that engagement of cognitive control optimizes a cost-benefit trade-off. In this work, we rely on Markov Decision Processes to derive the optimal control allocation policy in standard decision-making tasks, under arbitrary default preferences. Our working hypothesis is that decision confidence serves as the benefit term of this allocation problem, hence the “metacognitive” nature of decision control. Importantly, we provide behavioural evidence that the ensuing model accurately predicts whether and how mice override their default policy in the context of a repeated perceptual decision-making task. This opens an alleyway for assessing the brain circuits that operate the arbitration between default and controlled decision processes, in rodent models of both healthy and neuropsychiatric conditions.

## INTRODUCTION

What makes you ponder some decisions, but not others? Decisions permeate every aspect of our lives - what to eat, where to live, etc. - but the amount of effort that we put into different decisions varies tremendously. Rather than processing all decision-relevant information, we often rely on fast habitual and/or intuitive decision policies, which can lead to irrational biases and errors *(Kahneman et al., 1982)*. For example, snap judgments about others are prone to unconscious stereotyping, which often has enduring and detrimental consequences *(Greenwald & Banaji, 1995)*. Yet, we do not always follow the fast and negligent lead of habits or intuitions. So what determines how much time and effort we invest when making decisions?

Heedless decisions can be triggered by psychobiological determinants such as stress *(Porcelli et al., 2012; Porcelli & Delgado, 2009)*, emotions *(Harlé & Sanfey, 2007; Martino et al., 2006; Sokol-Hessner et al., 2013)* or fatigue *(Blain et al., 2016; Wiehler et al., 2022)*. But in fact, they also arise in the absence of such contextual factors. This is why they are sometimes viewed as the outcome of inherent neurocognitive constraints on the brain's decision processes, e.g., limited attentional and/or mnemonic capacity *(Giguère & Love, 2013; Lim et al., 2011; Marois & Ivanoff, 2005)* or unreliable neural representations of decision-relevant information *(Drugowitsch et al., 2016; Wang & Busemeyer, 2016; Wyart & Koechlin, 2016)*. However, an alternative perspective is that the brain has a preference for efficiency over accuracy *(Thorngate, 1980)*. For example, when making perceptual or motor decisions, people frequently trade accuracy for speed, even when time constraints are not tight *(Heitz, 2014; Palmer et al., 2005)*.

Over the past decade, computational accounts of decision making have formalized a simple and elegant mechanistic idea that relates decision time and accuracy. They assert that choices are only triggered once the accumulation of "evidence" in favor of alternative choice options crosses a critical decision threshold *(Ratcliff et al., 1999; Usher & Mcclelland, 2001)*, which is modulated by external incentives *(Frank, 2006; Simen et al., 2006)*. In turn, response times increase when the reward attached to correct decisions increases. Further computational work demonstrated that accumulation-to-bound policies, whose threshold is optimally set in the aim of maximizing decision *efficiency* (i.e. the ratio between expected reward and decision time), behave similarly *(Ditterich, 2006; Drugowitsch et al., 2012)*. From a theoretical perspective, this implies that heedless decisions can be viewed as an evolutionary adaptation, in response to selective pressure that penalize slow decision processes *(Pirrone et al., 2014)*.

A possibility is that slow processing of decision-relevant information incurs an *opportunity cost*, i.e. slow decision makers would forego otherwise obtainable rewards *(Agrawal et al., 2020; Kurzban et al., 2013)*. But time may not be the only cost to pay when pondering decisions. Recall that most decision contexts trigger (quick and effortless) automatic behavioral tendencies, which originate from learned habits or intuitions *(Schneider & Shiffrin, 1977)*. In humans, overriding such default/automatic behavioral tendencies involves the aversive experience of mental effort *(Massar et al., 2018; Sullivan-Toole et al., 2019)*. The latter can be thought of as the subjective cost of engaging cognitive control to pursue goal-oriented behavior *(Birnberg & Shields, 1984; Bordalo et al., 2017; Otto et al., 2025)*. Importantly, it was

recently suggested that cognitive control may be strategically deployed to maximize the so-called *expected value of control* or EVC *(Musslick et al., 2015; Ozcimder et al., 2017; Shenhav et al., 2013)*. Here, the EVC discounts the benefits of allocating cognitive control with its ensuing effort cost. Under this view, the brain tends to avoid investing cognitive control, except when the benefit of overriding default/automatic responses overcompensates its cost *(Cavanagh et al., 2013; Dunn et al., 2017; Kool et al., 2010)*. In humans, this idea has proven fruitful for understanding the relationship between decision time, mental effort, performance and confidence reports in various cognitive tasks *(Bénon et al., 2024; Griffiths et al., 2015; Lieder et al., 2018)*.

Although such computational mechanism received empirical support mostly in humans, they make qualitative and quantitative predictions regarding features of overt decision behavior – such as choice accuracy, decision time and confidence – that can, in principle, be observed in any animal species *(Bénon et al., 2024)*. The critical issue is whether a given species is equipped with the necessary cognitive apparatus, which rests on the ability to flexibly deploy cognitive control and basic metacognitive monitoring skills. In non-human primates, there is clear evidence that various monkey species can strategically mobilize cognitive control *(Brown & Hampton, 2020; De Petrillo et al., 2022)* and perform metacognitive evaluations over controlled policies *(Hampton, 2019; Smith et al., 2018; Subias et al., 2025)*. Though comparatively scarcer, similar evidence also exists in rodents. On the one hand, the metacognitive skills of rats *(Joo et al., 2021; Stolyarova et al., 2019)* and mice *(Findling et al., 2025; Kepecs et al., 2008)* are already established. On the other hand, despite its relevance for preclinical models of neuropsychiatric disorders *(Johnson et al., 2013; Robble et al., 2021)*, evidence for rodents' ability to flexibly override default/habitual responses through cognitive control is still debated, at least in mice *(Evans et al., 2019; Park et al., 2025)*. For example, even after extensive training on simple perceptual decision-making tasks, mice exhibit unequal performances from trial to trial, occasionally returning to suboptimal perseverative or heuristic (e.g. Win-Stay/Loose-Shift) policies *(Ashwood et al., 2022; Roy et al., 2021; Urai, 2026)*. Interestingly however, such within-individual behavioral variability clearly implies that mice do succeed in overriding default policies under some circumstances. Our working assumption is that rodents adapt their investment of cognitive control according to a cost-benefit tradeoff that varies across trials, following changes in, e.g., the relative value of default and controlled policies.

In humans, cognitive control is typically evidenced using conflicts between default/automatic and goal-oriented policies, which induce known interference effects in both performance and response time *(Cohen et al., 1990; Kalanthroff et al., 2018)*. What is lacking here is a quantitative framework for predicting the impact of control allocation on objective behavioral readouts (e.g., response time and performance), in cases where the default and instrumental responses are either congruent or incongruent, and as a function of motivational factors (such as reward at stake and task difficulty).
In this context, the contribution of this work is twofold. First, we extend previous computational work on EVC in decision making *(Bénon et al., 2024; Lee & Daunizeau, 2021)*, which essentially frames the allocation of cognitive control as an optimal stopping problem that trades decision confidence against decision time. Similarly, we rely on Markov decision processes to derive

the optimal stopping strategy under possibly conflicting default and controlled policies, and as a function of reward at stake and task difficulty. As we will see, this model predicts conflict-induced interference effects in both performance and response time, as well as their modulation by motivational factors. Second, we test whether the model can explain whether and how mice override default responses. This is important, because (transgenic lines of) mice are the most frequent rodent model for translational neuropsychiatric approaches. This paves the way for chemo- and opto-genetic investigations of the causal role of prefrontal subsystems in mental effort regulation. To do this, we developed a self-administered, automated, operant conditioning paradigm. During several weeks, mice stay in a customized conditioning box, in which they obtain food only when they successfully engage with a deterministic, binary, visual decision task. Critically, we systematically varied decision difficulty (i.e. visual cue discriminability) and reward (i.e. amount of food), and measure the ensuing performance and response time. This enabled us to track, over multiple days and in a pseudo-ecological environment, how mice regulate their effort, beyond variations in endogenous psycho-biological states (such as circadian cycle phase, hunger state, stress, etc) that may otherwise confound decision mechanisms.

## MATERIALS AND METHODS

### Animals

All experimental procedures followed national and European guidelines and were approved by the institutional review boards (French Ministry of Higher Education, Research and Innovation; APAFiS Protocols no. 1418-2015120217347265 and 2021042017105235). Animals were bred and group-housed in the animal facilities of the Paris Brain Institute in Tecniplast ventilated polycarbonate cages under positive pressure with hard-wood bedding in groups of up to six animals per cage with ab libitum food and water access. The temperature was maintained at 21–23 °C and the relative humidity at 55 ± 10\% with a 12-h light/dark cycle (lights on/off at 8am and 8pm, respectively). The animals were maintained in a 12-hour light/dark cycle (lights on/off at 8:00am/8:00 pm, respectively), and had *ab libitum* food and water access. In this study, n = 23 adult male mice were tested and no data excluded.

### Experimental setup

Operant conditioning was conducted in custom-modified experimental chambers (ENV-007CTX, Med Associates, Vermont, USA) as previously described (Benzina et al. 2021), in which each of the tested animals lives and performs the conditioning task 24/7 in an automated, self-initiated, self-paced manner. Seven such experimental operant chambers were operated in parallel in the same experimental room. Briefly, the rear wall of each chamber housed the feeder compartment (equipped with an infrared light beam to detect feeder entry crossing) and drinking water bottle holder, the front wall held two tactile screens, and a custom-adapted, homemade pair of black Plexiglas gates equipped with a pair of infrared (IR) beams. These beams allowed for tracking the locomotion of the mouse into or out of the area with the tactile screens. Precision reward pellets (20 mg LabTabTM AIN-76A rodent precision pellets, TestDiet, Richmond, USA) served as the sole nutrition during the entire experiment and all rewards had to be earned through successful performance of the task.

### Experimental task

Once placed in the operant conditioning chambers, the mice start with a pre- training phase of five different stages. In a first stage, a total of five precision pellets drop automatically once every minute if the previous pellet had successfully been recovered. In a second stage, the mouse is required to nose poke into the feeder compartment to trigger individual pellet delivery once per minute until a total of ten pellets has been received. In a third stage, crossing of the IR photobeams prompts the illumination of both tactile screens; the mouse is required to touch one of them to trigger pellet delivery in the feeder. There was no time limit for recovering the pellets at this stage. However, the condition for the mice to pass into the following pre-training stage was the recovery of the pellets within less than 15 seconds for at least 66% of the times. Then, in the consecutive pre- training stage, as before, both screens were illuminated with white light upon gate crossing. However, mice now had to nose poke into the feeder in less than 15 seconds after screen touch in order to trigger pellet delivery. Having successfully recovered at least 66% of the pellets in this manner, mice passed into the actual task. The last (fifth) pre-training stage consisted in a deterministic, binary visual discrimination task with two stimuli pairs. The presented stimulus pair consisted in ten, either

vertical or horizontal, white bars on a black screen, each of which was presented in a random manner (with an overall 50% frequency). To earn a precision pellet, mice need to associate each of the two visual cues with one screen touch, e.g. stimulus of ten horizontal/ vertical bars associated with the action of left/ right screen touch, accordingly. The cue was shown simultaneously on both touchscreens for two reasons: first, even if mice would only partially sample from one screen, they would have all information at their disposal to successfully solve the task; and, second, the mouse signals her decision (which screen she touches) independently from the way we present the visual cue. The cue-outcome contingencies were randomized across individuals and operant chambers. All trials were self-initiated and onsets of the visual cue pair on the tactile screens were triggered by the mouse crossing custom-designed, homemade photobeam-equipped gates. Mice were allowed a maximum of 60 seconds after stimulus onset to respond with a screen touch. Upon correct response, the screens started blinking to give a positive feedback for the mouse, which had 15 seconds to nose poke in the food dispenser in the rear wall for obtaining a food reward. Past 15 seconds after correct response, the trial became unrewarded. In case of incorrect screen touch, mice received an aversive white light for five seconds. Additionally, in order to decrease the tendency of mice to lateralize, each trial following an incorrect response required the mouse to switch to the other screen (so-called “corrective trials”). Past 60 seconds in the chamber compartment with the screens, the screens would switch off again; in these “time out” trials, mice then had to cross out and back into the feeder compartment to trigger the next trial. Trials where mice crossed in and out of the screen compartment within 60 seconds and without screen touch, were labelled as “aborted trials”.

Having reached 80% correct responses over the past 40 trials, the mice transitioned into the main (decision making) task, where we systematically modulated reward amounts and difficulty levels. Upon a correct response, the mouse earned either one or three reward pellets, under low or high reward condition, respectively. Mice subsequently performed low and high reward trial blocks of each 60 trials, consisting of 10 trials per difficulty level. Difficulty levels consisted of variations of the visual cue, whose association with the outcome was previously learned during pre-training stage five. More precisely, these variations consisted of different ratios of intersecting vertical and horizontal bars: the easiest difficulty level consisted of the same 10 vertical or 10 horizontal bars as in pre-training stage five; the next difficulty level consisted of 9:1 intersecting vertical and horizontal bars, etc, until five and five intersecting horizontal and vertical bars, representing the highest and unsolvable difficulty level, which was rewarded randomly 50% on each side. Except for the control 5:5 difficulty level, the cue-outcome contingency was identical to that of pre-training stage five, having identified the side with the dominant orientation of bars.

Importantly, following an incorrect response, the subsequent trial was constrained such that the correct side was systematically opposite to the animal’s previous choice. In the rest of the article, we refer to these trials following an error as *corrective trials*.

This design was intended to prevent perseverative responding on a single side, a strategy that would otherwise allow mice to obtain reward on half of the trials without the need to learn the task structure. By enforcing correct responses to be opposite to the previous choice, corrective

trials thus promote learning of the cue–response contingencies that determine reward delivery.

Figure 1 below shows the operant chamber and associated task design, as well as summary statistics of mice behavior in the task.

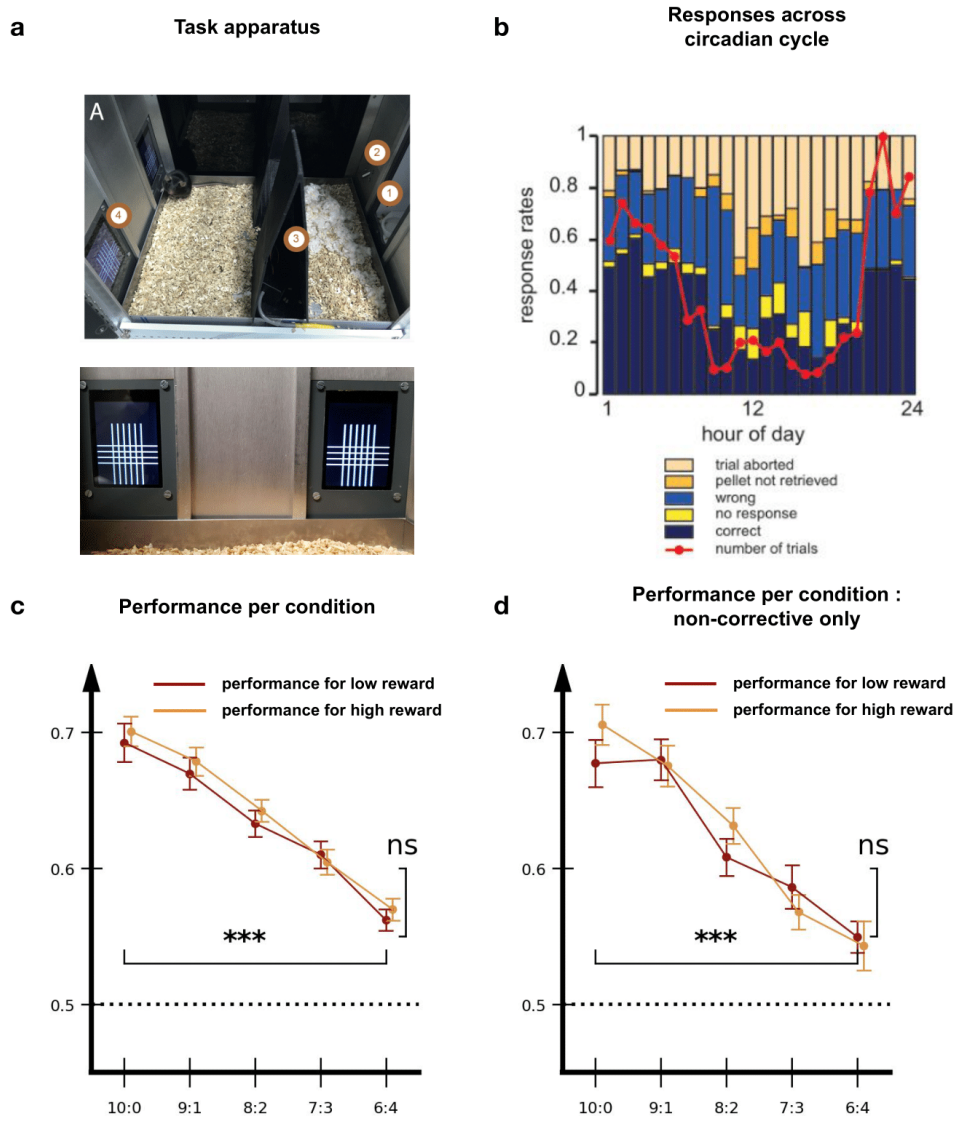


**Fig. 1 | Experimental design. (a)** Task apparatus. *Top:* Schematic of the experimental setup. (1) Food pellet dispenser; (2) lever used to obtain the reward; (3) entrance to the task zone, monitored by infrared detectors; and (4) tactile screens located within the task zone. *Bottom:* Close-up of the task zone showing the two symmetrical tactile screens positioned on the left and right sides. **(b)** Response distribution across the circadian cycle. | **(c)** Performance per difficulty in the main phase, with trial separated between low-reward (dark red) and high-reward (orange) conditions. Error bars indicate ±1 SEM. **(d)** Performance per difficulty in the main phase, computed on non-corrective trials only, with trial separated between low-reward (dark red) and high-reward (orange) conditions. Error bars indicate ±1 SEM.

In the remainder of this paper, we focus on correct and incorrect trials from the main phase of the experiment that followed a correct trial, that is, valid non-corrective trials. This effectively remove trials in which the cue can be predicted from the previous chosen action (corrective trials). Note that other trial types, e.g. aborted or unrewarded (about 22±7% of all non-corrective trials) were excluded.

**Computational model**

As we will see, mice do occasionally engage in some form of automatic/default policy, which effectively reduces to reinforcement learning. Although their default policy will occasionally yield relative successes (cf. corrective trials), mice have also learned the value of paying attention to the visual cue (cf. training phase). In principle, mice need to solve a credit assignment problem: what determines action outcomes? We assume that mice may be uncertain as to whether reward follows from selecting a given action (direct reward contingency, as is implicit in reinforcement learning), or from properly orienting action according to the cue (indirect reward contingency). This distinction is important because the former scenario only requires trial-by-trial monitoring of action-outcome contingencies, whereas the latter scenario induces a demand for processing the cue. In what follows, $\lambda$ denotes the mice's subjective estimate of the probability that actions are directly rewarded. For example, if $\lambda = 0$ , mice fully attribute rewards to properly orienting actions according to the cue, and therefore do not rely at all on their default policy.

Now, our computational model relies upon three additional assumptions:

(i) prior to processing the visual cue, mice possess a default/automatic action preference that derives from trial-by-trial reinforcement learning of action-outcome contingencies;

(ii) mice may override their default preference by engaging controlled processing of the visual cue, and

(iii) mice regulate their control investment on a decision-by-decision basis, by balancing the expected benefit of choosing the best action against the subjective cost incurred by spending time processing the cue.

Assumption (i) states that mice learn default values for candidate actions by updating a running average of the amount of reward that followed each candidate action. We denote these default action values by $Q^{\text{left}}_{\text{default}}$ and $Q^{\text{right}}_{\text{default}}$ . This yields a default/automatic preference, proportional to the difference in default action values $\Delta Q_{\text{default}}$. When the learning rate is high, this can be approximated by a very simple "win-stay/ lose-switch" heuristic policy, whereby the mouse tends to repeat (resp., switch from) a successful (resp. unsuccessful) action.

Assumption (ii) implies that, as mice spend more time sampling the visual cue, they increase the reliability of their perceptual evidence regarding the dominant orientation of the bars. It can be shown that, as mice randomly sample the visual scene, the (log-) ratio between vertical and horizontal visual energies, as computed from a Fourier decomposition of the sampled visual scene (Fig 2a, 2b), varies as a noisy signal $S_t$ centered on the relative orientation of the bars (Fig 2c).

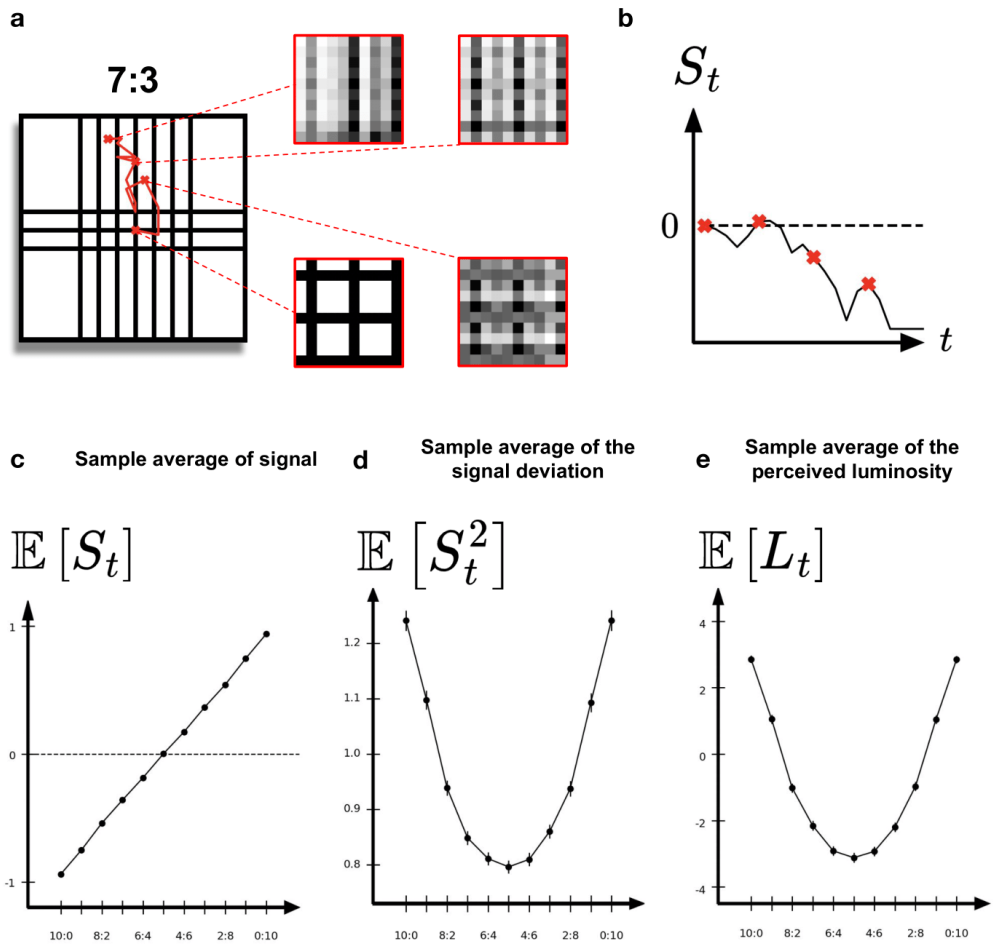


**Fig. 2 | Simulation of visual sampling of the stimulus by the mouse. (a)** Example trajectory of the modeled visual sampling process during a trial containing seven vertical and three horizontal lines (7:3 condition). The red path shows the sampling trajectory, modeled as a random walk over the visual scene. Red crosses mark successive sampling locations; the corresponding local visual patches perceived at those moments are displayed on the right. **(b)** Time course of the sampled signal along the visual trajectory shown in (a), corresponding to the log-ratio of sampled vertical and horizontal Fourier energy. Red crosses indicate time points corresponding to the sampled patches shown in (a). **(c)** Mean signal as a function of line contrast. Negative values indicate a predominance of vertical Fourier energy (as in the example shown in (a)), whereas positive values indicate a predominance of horizontal Fourier energy. **(d)** Mean signal squared deviation to zero as a function of line contrast. **(e)** Mean sampled stimulus luminosity (vertical plus horizontal Fourier log-energies) as a function of line contrast.

After having processed the cue up until time $t$, a simple Bayesian cue-denoising treatment of $S_{1:t}$ yields a time-dependent posterior belief about the dominant orientation of the bars:

$$\begin{cases} P(\text{cue says Left}|S_{1:t}) \simeq s\left(\dfrac{\pi\mu_t}{\sqrt{3}\sigma_t}\right) \\ P(\text{cue says Right}|S_{1:t}) = 1 - P(\text{cue says Left}|\mu_t, \sigma_t) \simeq s\left(\dfrac{-\pi\mu_t}{\sqrt{3}\sigma_t}\right) \end{cases}$$

where $s(\cdot)$ is the standard sigmoid mapping. Here, $\mu_t$ and $\sigma_t$ denote the signed evidence indicating which side is signaled by the cue and its associated uncertainty, respectively. The former is essentially a weighted average of past evidence signals $S_{1:t}$, and the latter follows exponentially decaying dynamics, starting at $\sigma_0$ (see Appendix). Importantly, the error rate of

decisions based upon the uncertain posterior belief about the true dominant orientation of the bars tends towards 0 as decision time tends to infinity.

Since the direction of cues are randomized, the Bayesian agent starts with a non-informative prior ($\mu_0 = 0$). Should a decision be based uniquely upon this prior belief, its error rate would be at chance level (50%). This also implies that the prior variance ${\sigma_0}^2$ reduces to the average squared deviation $E\left[S_t^2\right]$ of the signal to zero. It turns out that the latter is entirely driven trial difficulty (i.e., ambiguity of visual stimuli), which can be rapidly proxied by the average stimulus luminosity (Fig. 2d and 2e). This suggests that mice may prospectively respond to decision difficulty by adapting their prior variance at the onset of each decision trial.

Assumption (iii) requires mice to anticipate the costs and benefits of investing more cognitive control to process the cue. We model the former using a simple power function of decision time. The latter reduces to the expected reward $E[R|S_{1:t}]$, which depends on the reward at stake $R$, the internal monitoring of perceptual evidence, and the default action values (see Appendix):

$$E[R|S_{1:t}] = \max\{(1-\lambda)R \times P(\text{cue says Left}|S_{1:t}) + \lambda Q_{\text{default}}^{\text{Left}}, (1-\lambda)R \times P(\text{cue says Right}|S_{1:t}) + \lambda Q_{\text{default}}^{\text{Right}}\}$$

At trial onset, a non-zero default preference ($\Delta Q_{\text{default}} \neq 0$) induces an initial bias favoring one action (scaled by $\lambda$), despite absence of cue information ($\mu_0 = 0$). As perceptual evidence accumulates, action values are dynamically updated, and evidence in favor of the initially non-preferred action may override the default preference before commitment. Since processing the cue will generally increase the mice's confidence regarding the cue's dominant orientation, the incentive for investing more cognitive control will tend to increase with decision time. However, the degree to which increased confidence translates into higher net benefit depends on how strongly the mouse relies on her default policy ($\lambda$).

At any time $t$, an optimal agent will commit to a decision whenever its current net benefit is higher than the expected net benefit of continuing processing the cue. Both current and expected net benefits depend upon the evidence gathered so far, but the latter also requires anticipating future stop/continue decisions. Formally, this induces an optimal stopping problem, a specific instance of a Markov decision processes *(Papadimitriou & Tsitsiklis, 1987)* that can be solved using dynamic programming (see Appendix). In brief, the optimal stopping policy amounts to comparing $\mu_t$ to evidence boundaries that vary over decision time (and depend upon reward at stake and decision difficulty). Importantly, non-zero default preferences render these stopping boundaries asymmetric, such that less evidence is required to stop in favor of the default action. The main computational steps that are involved when deriving these boundaries are summarized in Figure 3 below.

Note that, over a broad region of parameter space, the model exhibits characteristic decision boundary dynamics, which effectively collapse as a function of elapsed decision time. In other

words, the longer the system spends processing the cue, the lower the amount of accumulated evidence required to terminate controlled processing and commit to a response.

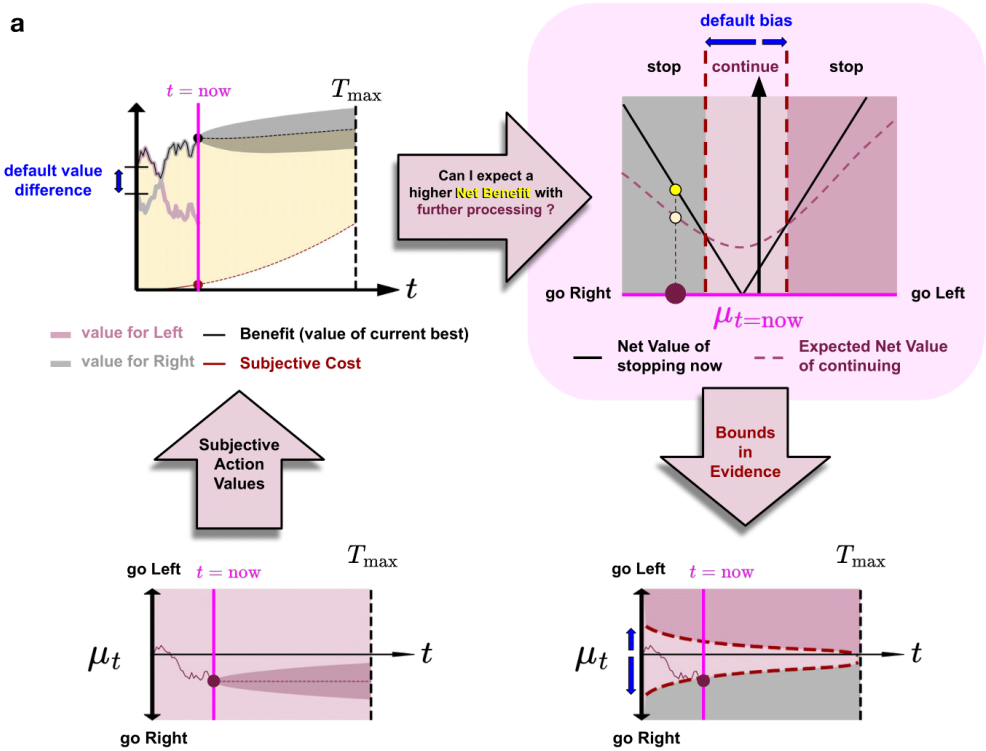


**Fig. 3 | Optimal stopping policy.** *Bottom left:* While processing the visual cue, the mouse accumulates noisy sensory evidence $\mu_t$ in favor of the direction indicated by the cue (purple trajectory). At a given moment in time (illustrated by the magenta vertical line), the system must decide whether to continue accumulating evidence or to commit to a choice. The dotted purple line and the shaded purple area represent, respectively, the expected mean and variance of the additional evidence that could be obtained by continuing to sample the cue. The maximum duration of the trial is indicated by the black dashed vertical line. *Top left:* Evidence accumulation modulates the subjective values of alternative actions. The value of choosing left (light purple) and right (light gray) evolve as evidence is accumulated. The instantaneous benefit (black curve) corresponds to the value of the currently best-estimated action. The subjective cost of continued processing (red curve) increases monotonically with time. The blue arrow denotes the baseline difference in action values in the absence of sensory evidence, reflecting the bias induced by the default policy. The red dotted line indicates the expected increase in cost with additional processing, whereas the black dotted line and shaded region denote the mean and variance of the benefit term with additional processing. *Top right:* At any given time, deciding whether to stop or continue accumulating evidence corresponds to solving an optimal stopping problem. For each level of accumulated evidence (magenta x-axis), the system can compute the net value of stopping now (solid black curve) and the expected net benefit of further processing (pink dashed curve). Depending on the current evidence level, either stopping is optimal (grey and purple side regions), or continuing is optimal (light red central region). These regions are separated by critical levels of evidence beyond which it is optimal to stop (dashed dark red vertical lines). The blue arrow highlights an asymmetry in these decision boundaries arising from the difference in default action values. *Bottom right:* Iterating this comparison at every time step yields time-dependent stopping boundaries that specify whether further processing should be engaged or whether a decision should be made, as a function of the currently accumulated evidence (dashed dark red curves). As in the top-right panel, blue arrows indicate a bias in these evidence thresholds induced by the default policy.

## Simulation procedure

All model simulations reported below were generated using a single parameterization of the model ($\sigma_L = 5$, $\alpha_C = 0.5$, $\nu_c = 2$, $\lambda = 0.05$ and $T_{\max} = 100$; see Appendix), with reward varied in the range $R \in [1.4; 2.6]$ and difficulty varied in the range $\sigma_0 \in [0.4; 0.6]$. In other

words, all predictions were jointly derived from a unique point in the parameter space. For each of the 10 reward × difficulty conditions (excluding the ambiguous 5:5 difficulty level), optimal decision bounds were computed, and 50,000 evidence trajectories were simulated under a Bayesian cue-denoising scheme. For each simulated trajectory, the correct side and the default side were drawn randomly and independently.

## RESULTS

### Characterization of response policies

We first assessed (i) whether mice correctly learned the cue-outcome contingency, and (ii) nonetheless partially relied on known default cue-independent response policies. Figure 4a summarizes the two response policies examined throughout the remainder of the study: controlled/instrumental (cue-following) and default/automatic (win-stay) responding.

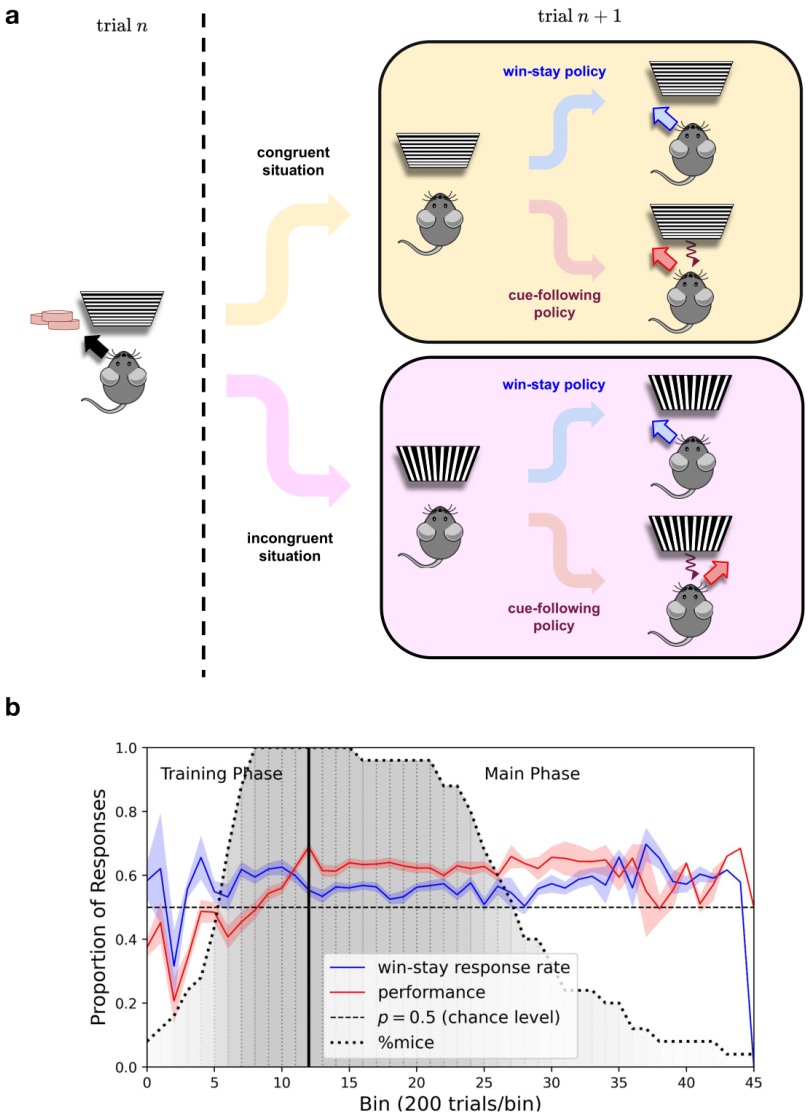


**Fig. 4 | Default and goal-directed policies in mice. (a)** Previous work shows that mice rely on at least two behavioral policies in such decision-making tasks. Here, we focus on non-corrective trials, which are those trials that followed a correct response (in the illustrated example, a mostly horizontal contrast is associated with a left response being correct at trial *n*). At trial *n*+1, the cue can point towards the same correct side as the previous trial (*congruent situation*, upper scenario, in yellow), or towards the opposite side (*incongruent situation*, lower scenario, in magenta). In any case, the mice can decide to follow its default policy (*win-stay*: blue arrow; left response in the illustrated example), or a controlled/cue-oriented policy (red arrow; corresponding to a left response in the congruent situation, and to a right response in the incongruent situation). By definition, a cue-oriented action is correct. **(b)** Dynamics of default (blue) and cue-following (red) policies across the final training phase and the main experimental phase. Each time bin contains 200 trials. Solid lines show the mean response rate and shaded areas indicate standard error of the mean (across animals). Data are aligned w.r.t. the onset of the main

experimental phase (thick vertical black line). The histogram reports the proportion of animals contributing to data within each time bin, reflecting variability across mice in the duration of training and in the total number of trials completed during the main phase.

Figure 4b shows how strongly mice rely on each of these two policies, across the final (fifth) training phase and the main phase of the experiment (analyses restricted to non-corrective valid trials). During the training phase, mice progressively improved their performance (red curve) while reducing their reliance on the initially dominating win-stay policy (blue curve), eventually reaching the 66% accuracy criterion required to enter the main phase. During the main phase, both performance and win-stay rate remained stable. Importantly, both the cue-following and win-stay response rates remained significantly above chance (one-sample *t*-test, $p < 0.001$). This implies that, although mice mostly rely on cues to orient their decisions, the default policy continued to influence behavior throughout the main phase.

**Evidence for cognitive control involvement: interference effects**

In humans, cognitive control skills are typically assessed using the so-called Stroop task, whereby a prepotent or default response tendency (e.g., word reading) interferes with a task-relevant or instrumental response (e.g., color naming). In this context, the ability to override the default policy in incongruent trials (i.e. when response policies conflict with each other) is interpreted as resulting from the engagement of cognitive control, which shifts the speed-accuracy tradeoff towards slower decisions *(Cohen et al., 1990; Kalanthroff et al., 2018)*.

Under the model, as cue information is processed within a given decision trial, evidence for the dominant bar orientation progressively accumulates and eventually converges towards the correct cue-indicated side in finite time (Fig. 5a, left panel). However, the optimal stopping policy imposes bounds on this accumulation process, such that not all evidence trajectories will have enough time to converge towards the correct cue side.

In a *congruent* situation (i.e., when the response side indicated by the visual cue matches the side favored by the win–stay default policy) correct responses require less evidence to trigger a decision than incorrect responses. This asymmetry facilitates correct choices and makes erroneous responses less likely (Fig. 5a, middle panel). The situation is reversed in *incongruent* situations (i.e., when the cue-signaled response conflicts with the side favored by the default policy ; Fig. 5a, right panel). This has two main consequences. First, incongruent trials typically yield lower performance than congruent trials, because a subset of noisy evidence trajectories that would have eventually converged towards the correct cue-indicated response in congruent trials prematurely reach the default boundary in incongruent trials (Fig. 5b, upper panel). Second, decisions in incongruent trials are slower than in congruent trials, because, although most trajectories are quickly attracted towards the correct cue-indicated side, the corresponding evidence boundary is further away in incongruent trials (Fig. 5c, upper panel).

We reasoned that providing behavioral evidence for the above interference effects could serve as a proof of principle that mice engage cognitive control to override their default policy. To test model predictions quantitatively, we regressed response side and response time for each

mouse against the side signaled by the cue, the side specified by the win–stay default policy, and their interaction. Regression coefficients were then analyzed at the group level using one-sample signed t-tests. Consistent with the model, we observed a significant effect of cue side ($p < 0.001$) as well as a significant effect of the default-specified side ($p < 0.001$) on response choice, with default preferences reducing the probability of responding to the cue-indicated side in incongruent trials (Fig. 5b, lower panel). In addition, we found a significant interaction between cue side and default side on response times ($p = 0.032$), reflecting longer response latencies in incongruent than in congruent conditions (Fig. 5c, lower panel).

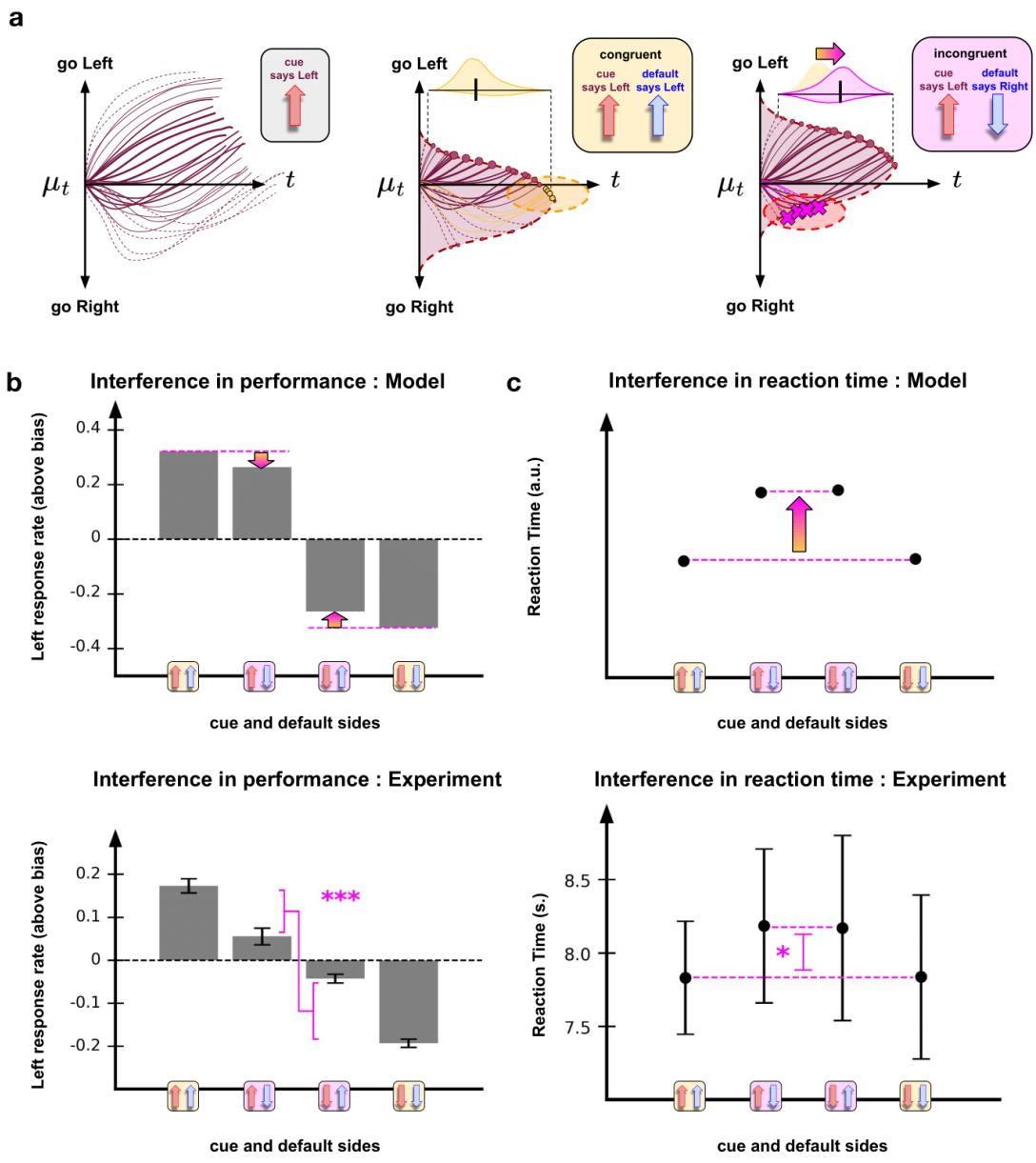


**Fig. 5 | Interference between default response mode and cue processing. (a)** Schematic illustration of evidence accumulation trajectories and decision boundaries under congruent and incongruent conditions. *Left:* When the cue specifies a left response, noisy sensory evidence favoring the left option is accumulated during cue processing. Because evidence accumulation is stochastic, individual trajectories do not immediately converge toward the correct response, but tend to do so given sufficient processing time. Trajectories differ in their likelihood, as indicated by line thickness. *Middle:* Optimal decision bounds truncate trajectories and induce a distribution of response times and response outcomes. When the system's default response is congruent with the cue (default = left), less evidence is

required to commit to the correct response than for the incorrect response. *Right:* When the default response is incongruent with the cue (default = right), this asymmetry is reversed: more evidence is required to reach the correct bound than for the incorrect (default) bound. This has two consequences. First, some trajectories that would have converged to the correct cue-consistent response under a congruent situation now terminate prematurely at the default bound, yielding incorrect responses (circled trajectories; endpoints shown as yellow disks in the middle panel and magenta crosses in the right panel). Second, in regimes where most trajectories still reach the correct bound, the expansion of the correct bound prolongs decision times for the majority of trials, resulting in an overall slowing of response times, illustrated by the rightward shift of the response-time distribution (arrow). **(b)** Testing predicted interference effects on performance. *Top:* model predictions (interference indicated by arrows). *Bottom:* experimental data. Error bars indicate ±1 SEM. **(c)** Testing predicted interference effects on response time. *Top:* model predictions (interference indicated by arrow). *Bottom:* experimental data. Error bars indicate ±1 SEM.

Orthogonally to these Stroop-like interference effects, the model also makes specific predictions regarding response latencies as a function of correctness and default-following. First, as in standard drift–diffusion models, it predicts longer response times for error trials than for correct trials, because most evidence accumulation trajectories are quickly attracted towards the correct decision boundary. For each mouse, we thus measure the mean response time difference between correct and incorrect trials, and test for a significant difference at the group level. This revealed a significant "slow error" effect ($p = 0.009$), as predicted by the model. Second, the model predicts longer response times for responses that do *not* follow the default (win–stay) policy than for those that do. This follows from the asymmetric decision bounds: less evidence is required to commit to a default-consistent response, all else being equal. Accordingly, we found that the average difference in response times between trials where mice either do or do not follow their default policy is significant at the group-level effect ($p = 0.023$).

**Motivational modulations of interference effects**

Having established that mice exhibit interference effects consistent with an arbitration between a default response mode and a controlled, cue-driven mode of responding, we next asked whether this arbitration is modulated by motivational factors that alter the cost–benefit balance of engaging controlled processing.

The model makes clear predictions regarding the impact of motivation on interference effects. Specifically, moving from a high-reward or low-difficulty condition to a low-reward or high-difficulty condition shifts the balance between the expected benefits and costs of allocating control : Reducing the reward at stake decreases the expected benefit of engaging control to process cue-related information. Similarly, increasing task difficulty increases the duration of controlled processing required to reach a given level of evidence, as lower stimulus contrast slows evidence accumulation. In both cases, the cost incurred to reach a given level of expected reward is increased. Thus, a system that seeks to maximize net benefit is ought to require less evidence before committing a decision. Accordingly, optimal stopping boundaries shrink when reward at stake decrease or decision difficulty increase (Fig. 6a, left to right).

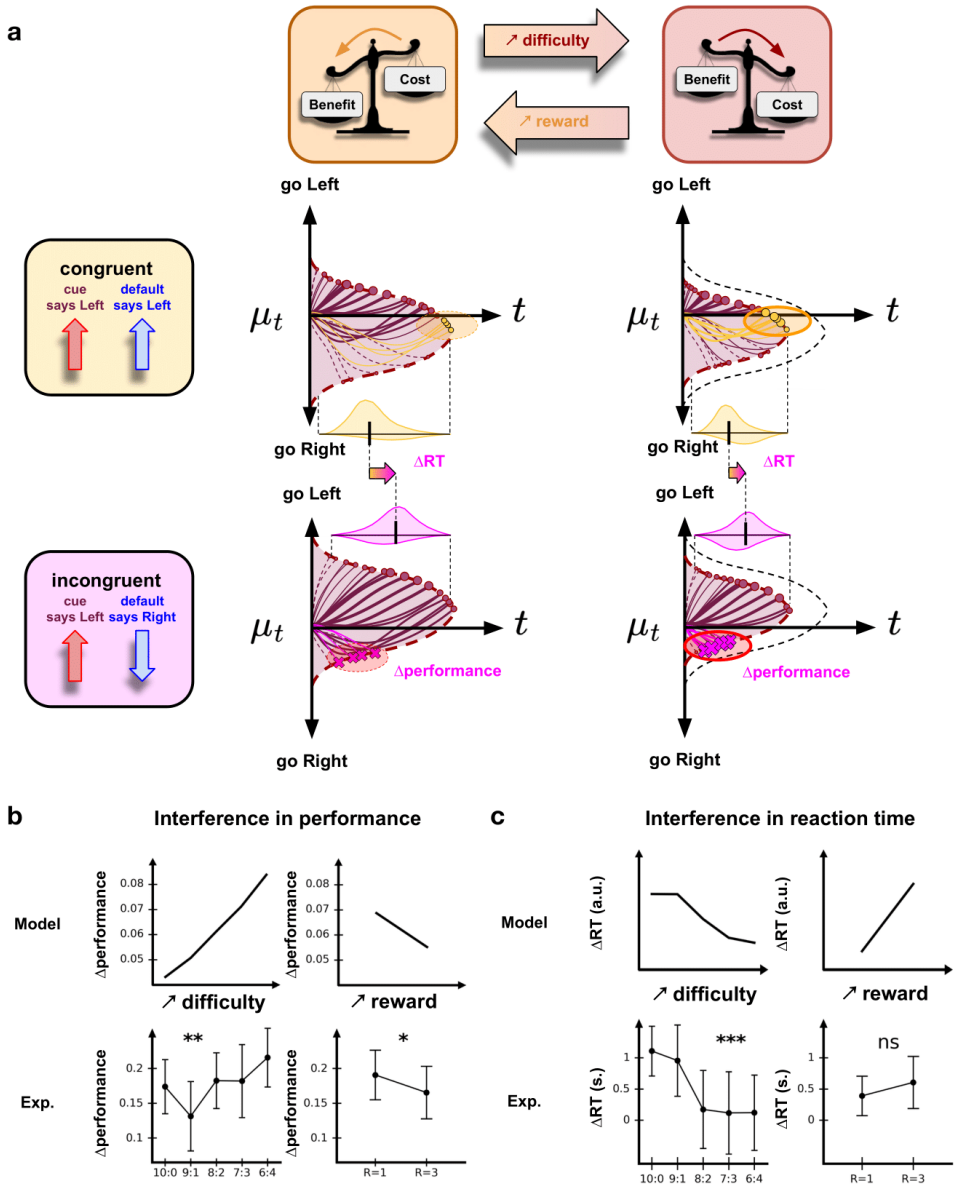


**Fig. 6 | Motivational modulation of interference effects. (a)** Schematic of evidence accumulation bounds under different motivational conditions. The upper row illustrates a congruent situation (cue = left, default = left), and the lower row an incongruent situation (cue = left, default = right). The left column corresponds to high-reward or low-difficulty conditions, whereas the right column corresponds to low-reward or high-difficulty conditions. Interference effects on response time (ΔRT) are indicated by arrows connecting the means of the congruent and incongruent response-time distributions within each motivational condition. Interference effects on performance (Δperformance) are illustrated by trajectories that are correct in the congruent condition (yellow disks) but become incorrect in the incongruent condition (magenta crosses). Decreasing the expected reward or increasing task difficulty shifts the benefit–cost trade-off of evidence accumulation, leading to a shrinkage of optimal decision boundaries (right column; high-reward/low-difficulty bounds are redrawn in dashed black lines for comparison). This boundary shrinkage reduces overall response times (thereby attenuating interference effects on response time), and increases the relative influence of default action preferences on choice (thus resulting in stronger interference effects on performance). **(b)** Testing the opposite motivational modulations of interference effects on performance by difficulty and reward, respectively. *Top:* model predictions. *Bottom:* experimental data (error bars indicate 95% confidence intervals). **(c)** Testing the opposite motivational modulations of interference effects on response time by difficulty and reward, respectively. *Top:* model predictions. *Bottom:* experimental data.

This leads to two distinct predictions regarding motivational modulations of interference effects.

First, when the stopping criterion becomes more liberal, decisions are made on the basis of lower-quality evidence. As a result, action values are comparatively more influenced by their

default bias, meaning that initial default preferences are less likely to be revised. Consequently, the impact of the default policy on choice increases, leading to stronger interference effects on performance (i.e. larger differences in accuracy between congruent and incongruent trials) when reward decreases or difficulty increases (Fig 6b). We tested this prediction by regressing binary performance data against congruence, difficulty (or reward), and their interaction for each mouse. Group-level analyses of regression coefficients revealed a significant positive interaction between difficulty and performance interference ($p = 0.006$), as well as a significant negative interaction between reward and performance interference ($p = 0.037$), in line with model predictions.

Second, because the system requires less evidence overall, response times decrease in both congruent and incongruent trials. This global speeding tends to reduce the difference in response times between congruent and incongruent conditions. In other words, the model predicts that the above interference effect on response time should decrease as reward decreases or difficulty increases (Fig. 6c). To test this prediction, we regressed response times for each mouse against trial congruence (whether the cue signaled the same side as the default policy), difficulty (or reward at stake), and their interaction. Regression coefficients were then analyzed at the group level using one-sample signed t-tests. As predicted by the model, we observed a significant negative interaction between difficulty and response-time interference ($p < 0.001$), indicating reduced interference at higher difficulty. However, we didn't find any significant interaction between reward and response-time interference ($p = 0.233$).

**Motivational modulation on the arbitration between response modes**

The model also predicts that motivational factors will directly modulate the relative engagement of default and cue-driven response modes. Recall that increasing trial difficulty or decreasing reward at stake will shift the balance between the benefits and the costs of control, such that less controlled processing is required to commit to a decision. In both cases, the ensuing shrinkage of the bounds has three direct effects on the observed responses (see Fig 7 below).

First, the global shortening of controlled processing time holds, irrespective of whether responses eventually follow the cue and/or the default policy. To test this prediction, we regressed response times for each mouse against trial difficulty and reward at stake, having splat trials according to either correctness (correct versus incorrect) or default-following (win-stay versus win-switch). Regression coefficients were then analyzed at the group level using one-sample signed t-tests. As predicted, we observed a significant positive effect of reward and a significant negative effect of difficulty on overall response times ($p < 0.001$ and $p = 0.005$, respectively). Response times for both correct and incorrect trials showed the same pattern, with significant decreases with difficulty ($p = 0.018$ and $p = 0.004$) and significant increases with reward ($p < 0.001$ and $p < 0.001$). In addition, difficulty significantly reduced response times for win–switch responses ($p = 0.008$), although no significant effect was observed for win–stay responses ($p = 0.203$). Conversely, reward significantly increased response times for both win–stay and win–switch responses ($p < 0.001$ and $p < 0.001$, respectively).

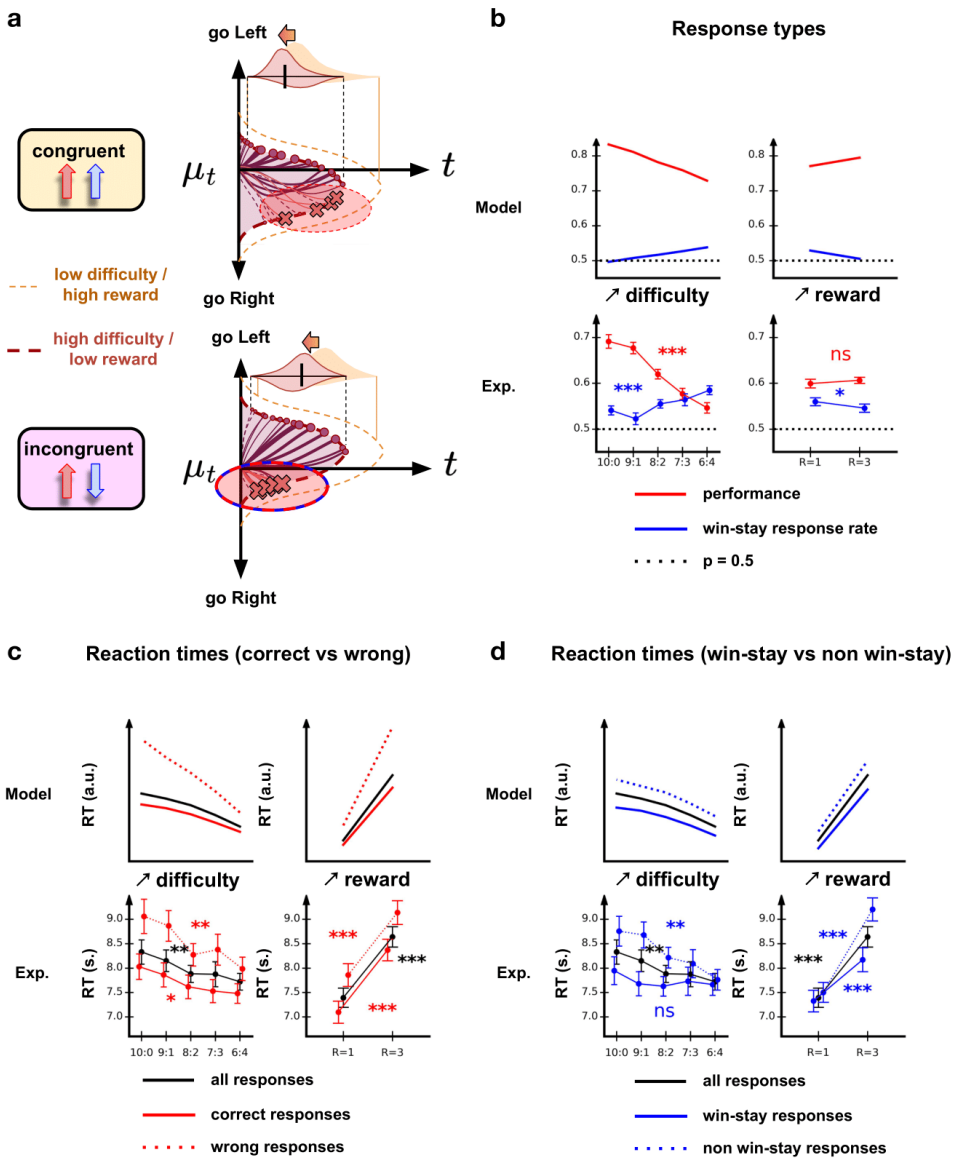


**Fig. 7 | Motivational modulation of arbitration between policies. (a)** Schematic of the model's evidence bounds. The upper row corresponds to congruent trials (cue = left, default = left), and the lower row to incongruent trials (cue = left, default = right). Gold dashed bounds denote the high-reward/low-difficulty condition, whereas dark red bounds denote the low-reward/high-difficulty condition. Decreasing reward or increasing difficulty shifts response time distributions toward shorter latencies. As a result, some trajectories that reach the correct bound in the high-reward/low-difficulty condition fail to do so in the low-reward/high-difficulty condition (red crosses). Critically, this effect is amplified in incongruent trials (blue and red dashed circle), where incorrect responses are consistent with the win-stay policy. Consequently, assuming an equal proportion of congruent and incongruent trials (as in non-corrective trials), the model predicts an increase in win-stay responses with increasing difficulty or decreasing reward. **(b)** The model predicts opposite motivational modulations of performance (red) and win-stay response rate (blue) as a function of reward and difficulty. Top: model predictions. Bottom: experimental data. **(c)** The model predicts opposite motivational modulations of response times for correct (solid red) and incorrect (dashed red) responses as a function of reward and difficulty. *Top*: model predictions. *Bottom*: experimental data. Error bars indicate ±1 SEM. **(d)** The model predicts opposite motivational modulations of response times for win-stay (solid blue) and non–win-stay (dashed blue) responses. *Top*: model predictions. *Bottom*: experimental data. Error bars indicate ±1 SEM.

Second, reducing controlled processing time limits the opportunity for initially erroneous evidence trajectories to eventually reach the correct decision boundary. In turn, a subset of trajectories that would have led to correct responses in high-reward or low-difficulty conditions are now terminated prematurely and result in errors. Thus, the model predicts that the rate of correct (cue-following) responses decreases when either difficulty increases or reward

decreases. Consistent with this prediction, increasing difficulty led to a significant decrease in performance ($p < 0.001$), in line with model's expectations. However, regressing performance against reward didn't reveal any significant effect ($p = 0.1825$).

Third, and for the same reason, reducing controlled processing time eventually promotes decisions that are still driven by default preferences. In other words, the model predicts a net increase in default-following responses when either difficulty increases or reward decreases. To test this prediction, we regressed the probability of emitting a win–stay response against reward and difficulty for each mouse. Group-level analyses revealed a significant negative effect of reward on win–stay rate ($p = 0.035$), as well as a significant positive effect of difficulty ($p < 0.001$), both in agreement with model predictions.

**Relationship between decision time and interference effects**

Finally, the model makes additional predictions regarding the relationship between response time on the one hand, and response rates (cue-following or default-following) or performance interference effects on the other hand (Fig. 8a).

To begin with, let us focus on the relationship between response time and response rates.

First, early responses are typically associated with trajectories in which strong evidence rapidly accumulated in favor of one option, whereas late responses correspond to flatter, less decisive evidence trajectories. As a result, late decisions are inherently more susceptible to noise, i.e. the model predicts that the rate of correct (cue-following) responses monotonically decreases when response time increases, even when controlling for motivational factors (Fig. 8b). Consistent with this prediction, when regressing performance against response time for each mouse and testing regression coefficients at the group level, we observed a significant negative effect of response time on performance ($p = 0.002$). Importantly, this relationship remained significant after regressing out the effects of task difficulty and reward ($p < 0.001$).

Second, response time also modulates the tendency to follow the default (win–stay) policy. Recall that the default decision boundary is closer to the starting state of evidence trajectories than the non-default decision boundary. Consequently, early in the trial, trajectories are more likely to terminate at the default boundary. As decision time elapses however, and trajectories that are favorable to the default response are progressively absorbed, the pool of remaining trajectories converges to a more neutral support of default versus non-default responses. In turn, the model predicts that late responses are less likely to follow the default policy, even when controlling for motivational factors (Fig. 8b). In line with this prediction, regressing win–stay responses against response time revealed a significant negative effect of response time on default-following behavior ($p < 0.001$), which again remained significant after controlling for reward and difficulty ($p < 0.001$).

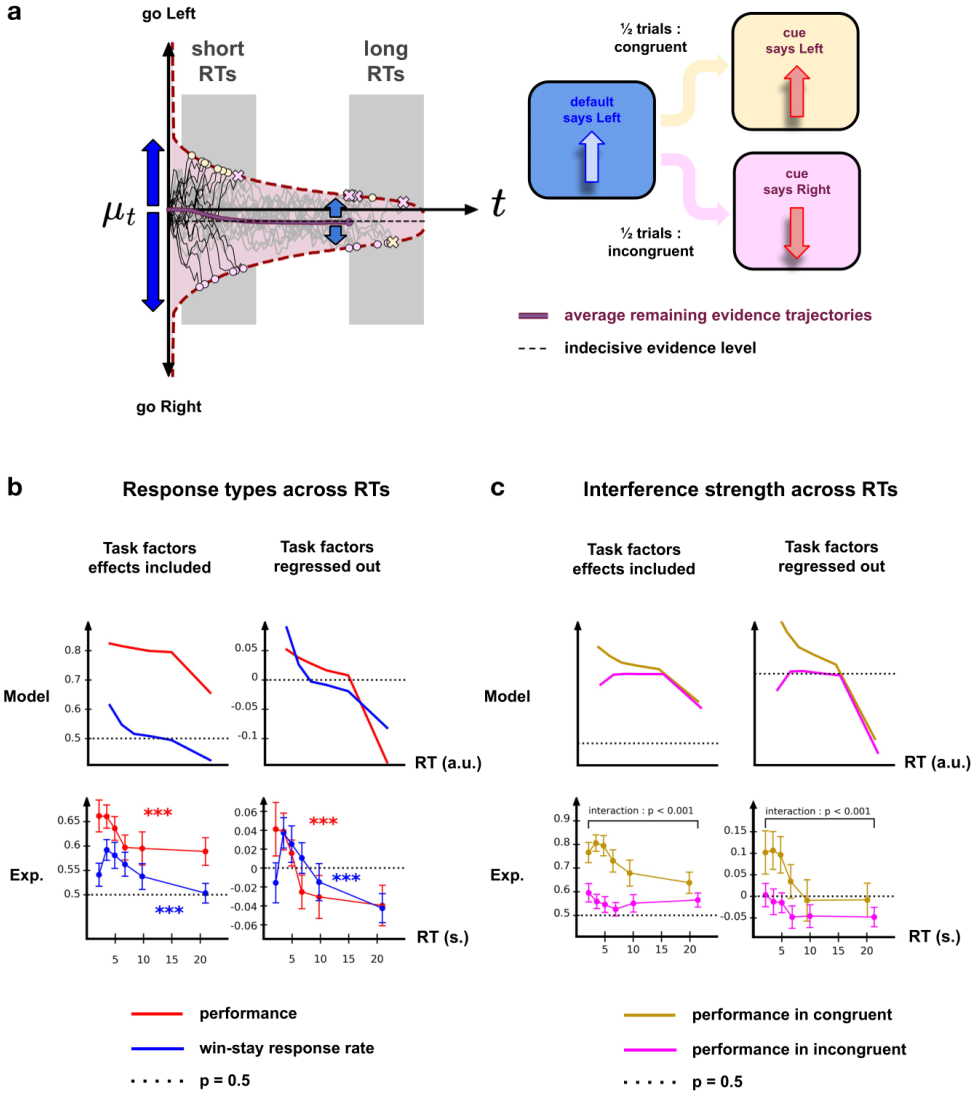


**Fig. 8 | Linking response time to arbitration between policies and interference.** | **(a)** Simulation of the model's evidence bounds under a fixed default policy favoring the left response. Half of the trials (yellow) correspond to congruent trials in which the cue indicates left, and the other half (magenta) to incongruent trials in which the cue indicates right. The two shaded grey regions on the evidence bounds represent short and long response-time intervals; corresponding evidence trajectories are shown as dark and light solid traces, respectively. Correct and incorrect responses are indicated by disks and crosses respectively. Trial congruency is encoded by the color of the trajectory endpoint (yellow for congruent trials, magenta for incongruent trials). The purple trace represents the average evidence trajectory over the set of remaining (non-terminated) trials as a function of elapsed time. The black dashed line denotes the indecisive evidence level, defined as the amount of evidence for which both actions lead to the same expected value. As the trial unfolds, the mean evidence of the remaining trajectories gets closer to this indecisive level. Dark and light blue arrows illustrate the average amount of extra evidence required to commit to a right or left decision, in the short and long response-time regions, respectively. | **(b)** The model predicts systematic variations of performance and win-stay rate as a function of response time. *Top*: model predictions. *Bottom*: experimental data. Error bars indicate 95% confidence intervals. The left column shows the relationship between response type and response time across task conditions (i.e. difficulty and reward), whereas the right column shows the same relationship after regressing out task conditions. | **(c)** The model predicts how interference effects on performance vary with response time. *Top*: model predictions. *Bottom*: experimental data. Error bars indicate 95% confidence intervals. The left column shows the relationship across task conditions (i.e. difficulty and reward), and the right column shows the relationship after regressing out task conditions.

A closer inspection of the above progressive selection of evidence trajectories reveals a non-trivial modulation of performance interference effects by response time.

Recall that it takes some controlled processing time until evidence trajectories filter the noise out and become progressively attracted towards the cue-following (correct) boundary.

Accordingly, the untruncated distribution of evidence trajectories tends to wrap around ballistic dynamics that progressively converge towards the cue-following boundary (cf. Fig 5a). Evidence noise broadens this distribution, such that unreliable evidence trajectories promote either premature hits of the closest (default) decision boundary or late hits of either boundary, independently of cue-related information. In congruent trials, hitting the default boundary yields a correct response. Thus, early hits are more likely to be correct than late hits, either because of default preferences (which facilitate premature though lucky boundary hits) or because of controlled processing of cue-related information. In fact, late hits correspond to those rare trajectories where evidence noise somehow compensated for cue-related information. In turn, the rate of correct responses monotonically decreases as decision time unfolds. In incongruent trials, premature hits driven by unreliable evidence rather compromise response correctness, until the distribution of the remaining evidence trajectories progressively truncates the tail that promotes early responses. In this case, the rate of correct responses should start low, gradually increases, and then falls again as decision increases and late hits correspond to trajectories that are comparable to those of congruent trials. The model thus predicts that the performance advantage of congruent trials is highest for quick response times, and eventually dampens as decision time increases. Note that this prediction holds even when controlling for motivational factors.

This prediction was tested by regressing performance against response time, trial congruence, and their interaction. Group-level analysis revealed a significant interaction between response time and trial congruence, indicating that interference effects on performance are indeed larger at shorter response times ($p < 0.001$). Crucially, this interaction persisted after regressing out the effects of reward and difficulty ($p < 0.001$).

## DISCUSSION

In this work, we have provided behavioral evidence that mice override their default policy by engaging cognitive control according to a motivational cost/benefit a tradeoff. On the computational side of things, we have shown whether and how motivational factors such reward at stake and decision difficulty eventually modify this tradeoff and thus impact the arbitration between default and controlled policies. On the behavioral side of things, we have validated all the ensuing model predictions, in the context of an experimental setup with high ecological validity. This opens an alleyway to assessing the brain circuits that operate the arbitration between default and controlled decision processes.

First, this work has implications for the neural architecture underlying the control of goal-directed behavior. Previous empirical studies relying on accumulation-to-bound models provided evidence that the adjustment of decision thresholds may be operated by the basal ganglia, in response to inputs from the orbitofrontal cortex or OFC *(Frank, 2006)*. This is consistent with the fact that disruption of subthalamic nucleus or striatal activity may cause premature responding in humans *(Frank et al., 2007)* and promote default -e.g. perseverant-responses in mice *(Burguière et al., 2013)*. In contrast, exhaustive neuroimaging meta-analyses indicated that cost/benefit computations involved in cognitive control very likely engage the anterior cingulate cortex or ACC *(Shenhav et al., 2013)*, which has known direct anatomical connections to the OFC and the basal ganglia *(Vogt & Paxinos, 2014)*. This raises the questions of whether and how ACC contributes to the control of decision making. We are currently addressing this issue using chemogenetic manipulations of ACC activity in mice.

Second, this work has implications for pathological neuropsychiatric conditions such as MDD[1], OCD[2] and ADHD[3], whose clinical symptoms can be understood in terms of specific forms of aberrant allocation of mental resources. For example, patients suffering from treatment-resistant depression typically exhibit variants of the so-called dysexecutive syndrome, i.e. co-occurring loss of cognitive/affective/behavioral control. Under our modelling framework, loss of control can have multiple causes, including: reduced sensitivity to reward, heightened sensitivity to mental effort costs, distorted beliefs regarding the relationship between one's effort and one's resulting cognitive performance. A priori, these causes map to distinct symptomatic dimensions, i.e. anhedonia, abulia, and low self-esteem, respectively. Thus, should our neural investigations of metacognitive decision control processes be successful, the dual computational/behavioral approach we present here may be fruitfully adapted to address outstanding issues in translational neuropsychiatry.

**REFERENCES**


Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2020). The Temporal Dynamics of Opportunity Costs : A Normative Account of Cognitive Fatigue and Boredom. *bioRxiv*, 2020.09.08.287276. https://doi.org/10.1101/2020.09.08.287276

Ashwood, Z. C., Roy, N. A., Stone, I. R., Urai, A. E., Churchland, A. K., Pouget, A., & Pillow, J. W. (2022). Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, *25*(2), 201-212. https://doi.org/10.1038/s41593-021-01007-z

Bénon, J., Lee, D., Hopper, W., Verdeil, M., Pessiglione, M., Vinckier, F., Bouret, S., Rouault, M., Lebouc, R., Pezzulo, G., Schreiweis, C., Burguière, E., & Daunizeau, J. (2024). The online metacognitive control of decisions. *Communications Psychology*, *2*(1), 1-17. https://doi.org/10.1038/s44271-024-00071-y

Birnberg, J. G., & Shields, M. D. (1984). The role of attention and memory in accounting decisions. *Accounting, Organizations and Society*, *9*(3), 365-382. https://doi.org/10.1016/0361-3682(84)90020-5

Blain, B., Hollard, G., & Pessiglione, M. (2016). Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of Sciences*, *113*(25), 6967-6972. https://doi.org/10.1073/pnas.1520527113

Bordalo, P., Gennaioli, N., & Shleifer, A. (2017). Memory, Attention, and Choice. *Quarterly Journal of Economics*.

Brown, E. K., & Hampton, R. R. (2020). Cognitive control of working memory but not familiarity in rhesus monkeys (Macaca mulatta). *Learning & Behavior*, *48*(4), 444-452. https://doi.org/10.3758/s13420-020-00432-7

Burguière, E., Monteiro, P., Feng, G., & Graybiel, A. M. (2013). Optogenetic Stimulation of Lateral Orbitofronto-Striatal Pathway Suppresses Compulsive Behaviors. *Science*, *340*(6137), 1243-1246. https://doi.org/10.1126/science.1232380

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal Theta Overrides Pavlovian Learning Biases. *Journal of Neuroscience*, *33*(19), 8541-8548. https://doi.org/10.1523/JNEUROSCI.5754-12.2013

Cohen, J. D., Mcclelland, J. L., & Dunbar, K. (1990). On the control of automatic processes : A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332-361.

De Petrillo, F., Nair, P., Cantwell, A., & Rosati, A. G. (2022). The Evolution of Cognitive Control in Lemurs. *Psychological Science*, *33*(9), 1408-1422. https://doi.org/10.1177/09567976221082938

Ditterich, J. (2006). Evidence for time-variant decision making. *The European Journal of Neuroscience*, *24*(12), 3628-3641. https://doi.org/10.1111/j.1460-9568.2006.05221.x

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience*, *32*(11), 3612-3628. https://doi.org/10.1523/JNEUROSCI.4010-11.2012

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, *92*(6), 1398-1411. https://doi.org/10.1016/j.neuron.2016.11.005

Dunn, T. L., Koehler, D. J., & Risko, E. F. (2017). Evaluating Effort : Influences of Evaluation Mode on Judgments of Task-specific Efforts. *Journal of Behavioral Decision Making*, *30*(4), 869-888. https://doi.org/10.1002/bdm.2018

Evans, D. A., Stempel, A. V., Vale, R., & Branco, T. (2019). Cognitive Control of Escape Behaviour. *Trends in Cognitive Sciences*, *23*(4), 334-348. https://doi.org/10.1016/j.tics.2019.01.012

Findling, C., Hubert, F., Acerbi, L., Benson, B., Benson, J., Birman, D., Bonacchi, N., Buchanan, E. K., Bruijns, S., Carandini, M., Catarino, J. A., Chapuis, G. A., Churchland, A. K., Dan, Y., Davatolhagh, F., DeWitt, E. E. J., Engel, T. A., Fabbri, M.,

Faulkner, M. A., … Pouget, A. (2025). Brain-wide representations of prior information in mouse decision-making. *Nature*, *645*(8079), 192-200. https://doi.org/10.1038/s41586-025-09226-1

Frank, M. J. (2006). Hold your horses : A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks: The Official Journal of the International Neural Network Society*, *19*(8), 1120-1136. https://doi.org/10.1016/j.neunet.2006.03.006

Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold Your Horses : Impulsivity, Deep Brain Stimulation, and Medication in Parkinsonism. *Science*, *318*(5854), 1309-1312. https://doi.org/10.1126/science.1146157

Giguère, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences*, *110*(19), 7613-7618. https://doi.org/10.1073/pnas.1219674110

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition : Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4-27.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources : Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217-229.

Hampton, R. R. (2019). Monkey Metacognition Could Generate More Insight. *Animal Behavior and Cognition*, *6*(4), 230-235. https://doi.org/10.26451/abc.06.04.02.2019

Harlé, K. M., & Sanfey, A. G. (2007). Incidental sadness biases social economic decisions in the Ultimatum Game. *Emotion (Washington, D.C.)*, *7*(4), 876-881. https://doi.org/10.1037/1528-3542.7.4.876

Heitz, R. P. (2014). The speed-accuracy tradeoff : History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00150

Johnson, A. W., Jaaro-Peled, H., Shahani, N., Sedlak, T. W., Zoubovsky, S., Burruss, D., Emiliani, F., Sawa, A., & Gallagher, M. (2013). Cognitive and motivational deficits together with prefrontal oxidative stress in a mouse model for neuropsychiatric illness. *Proceedings of the National Academy of Sciences*, *110*(30), 12462-12467. https://doi.org/10.1073/pnas.1307925110

Joo, H. R., Liang, H., Chung, J. E., Geaghan-Breiner, C., Fan, J. L., Nachman, B. P., Kepecs, A., & Frank, L. M. (2021). Rats use memory confidence to guide decisions. *Current Biology: CB*, *31*(20), 4571-4583.e4. https://doi.org/10.1016/j.cub.2021.08.013

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty : Heuristics and Biases*. Cambridge University Press.

Kalanthroff, E., Davelaar, E. J., Henik, A., Goldfarb, L., & Usher, M. (2018). Task conflict and proactive control : A computational theory of the Stroop task. *Psychological Review*, *125*(1), 59-82. https://doi.org/10.1037/rev0000083

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227-231. https://doi.org/10.1038/nature07200

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665-682. https://doi.org/10.1037/a0020198

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(06), 661-679. https://doi.org/10.1017/S0140525X12003196

Lee, D. G., & Daunizeau, J. (2021). Trading mental effort for confidence in the metacognitive control of value-based decision-making. *eLife*, *10*, e63282. https://doi.org/10.7554/eLife.63282

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLOS Computational Biology*, *14*(4), e1006043. https://doi.org/10.1371/journal.pcbi.1006043

Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2011). The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. *Journal of Neuroscience*, *31*(37), 13214-13223. https://doi.org/10.1523/JNEUROSCI.1246-11.2011

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, *9*(6), 296-305. https://doi.org/10.1016/j.tics.2005.04.010

Martino, B. D., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, Biases, and Rational Decision-Making in the Human Brain. *Science*, *313*(5787), 684-687. https://doi.org/10.1126/science.1128356

Massar, S. A. A., Csathó, Á., & Van der Linden, D. (2018). Quantifying the Motivational Effects of Cognitive Fatigue Through Effort-Based Decision Making. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00843

Musslick, S., Shenhav, A., Botvinick, M., & D Cohen, J. (2015, juin 7). *A Computational Model of Control Allocation based on the Expected Value of Control*. Conference: Reinforcement Learning and Decision Making 2015.

Otto, A. R., Westbrook, A., & Daunizeau, J. (2025). Why is cognitive effort experienced as costly? *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2025.10.012

Ozcimder, K., Dey, B., Musslick, S., Petri, G., Ahmed, N. K., Willke, T. L., & Cohen, J. D. (2017). A Formal Approach to Modeling the Cost of Cognitive Control. *arXiv:1706.00085 [q-bio]*. http://arxiv.org/abs/1706.00085

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 376-404. https://doi.org/10.1167/5.5.1

Papadimitriou, C. H., & Tsitsiklis, J. N. (1987). The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, *12*(3), 441-450. https://doi.org/10.1287/moor.12.3.441

Park, E. H., Sparks, K. C. O., Grubbs, G., Taborga, D., Nicholas, K., Ahmed, A. S., Ruiz-Pérez, N., Kim, N., Segura-Carrillo, S., & Fenton, A. A. (2025). Cognitive control of behavior and hippocampal information processing without medial prefrontal cortex. *eLife*, *13*. https://doi.org/10.7554/eLife.104475.2

Pirrone, A., Stafford, T., & Marshall, J. A. R. (2014). When natural selection should optimize speed-accuracy trade-offs. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00073

Porcelli, A. J., & Delgado, M. R. (2009). Acute stress modulates risk taking in financial decision making. *Psychological Science*, *20*(3), 278-283. https://doi.org/10.1111/j.1467-9280.2009.02288.x

Porcelli, A. J., Lewis, A. H., & Delgado, M. R. (2012). Acute Stress Influences Neural Circuits of Reward Processing. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00157

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261-300. https://doi.org/10.1037/0033-295X.106.2.261

Robble, M. A., Schroder, H. S., Kangas, B. D., Nickels, S., Breiger, M., Iturra-Mena, A. M., Perlo, S., Cardenas, E., Der-Avakian, A., Barnes, S. A., Leutgeb, S., Risbrough, V. B., Vitaliano, G., Bergman, J., Carlezon, W. A., & Pizzagalli, D. A. (2021). Concordant neurophysiological signatures of cognitive control in humans and rats. *Neuropsychopharmacology*, *46*(7), 1252-1262. https://doi.org/10.1038/s41386-021-00998-4

Roy, N. A., Bak, J. H., International Brain Laboratory, Akrami, A., Brody, C. D., & Pillow, J. W. (2021). Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron*, *109*(4), 597-610.e6. https://doi.org/10.1016/j.neuron.2020.12.004

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing : I. Detection, search, and attention. *Psychological Review*, *84*(1), 1-66. https://doi.org/10.1037/0033-295X.84.1.1

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The Expected Value of Control : An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, *79*(2), 217-240. https://doi.org/10.1016/j.neuron.2013.07.007

Simen, P., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, *19*(8), 1013-1026. https://doi.org/10.1016/j.neunet.2006.05.038

Smith, T., Smith, J., & Beran, M. (2018). Not knowing what one knows : A Meaningful failure of metacognition in capuchin monkeys. *Animal Behavior and Cognition*, *5*, 55-67. https://doi.org/10.26451/abc.05.01.05.2018

Sokol-Hessner, P., Camerer, C. F., & Phelps, E. A. (2013). Emotion regulation reduces loss aversion and decreases amygdala responses to losses. *Social Cognitive and Affective Neuroscience*, *8*(3), 341-350. https://doi.org/10.1093/scan/nss002

Stolyarova, A., Rakhshan, M., Hart, E. E., O'Dell, T. J., Peters, M. a. K., Lau, H., Soltani, A., & Izquierdo, A. (2019). Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nature Communications*, *10*(1), 4704. https://doi.org/10.1038/s41467-019-12725-1

Subias, L., Katsu, N., & Yamada, K. (2025). Metacognition in nonhuman primates : A review of current knowledge. *Primates; Journal of Primatology*, *66*(1), 9-25. https://doi.org/10.1007/s10329-024-01169-x

Sullivan-Toole, H., DePasque, S., Holt-Gosselin, B., & Galván, A. (2019). Worth working for : The influence of effort costs on teens' choices during a novel decision making game. *Developmental Cognitive Neuroscience*, *37*, 100652. https://doi.org/10.1016/j.dcn.2019.100652

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science*, *25*(3), 219-225. https://doi.org/10.1002/bs.3830250306

Urai, A. E. (2026). Structure uncovered : Understanding temporal variability in perceptual decision-making. *Trends in Cognitive Sciences*, *30*(1), 54-65. https://doi.org/10.1016/j.tics.2025.06.003

Usher, M., & Mcclelland, J. L. (2001). On the time course of perceptual choice : The leaky competing accumulator model. *Psychological Review*, 550-592.

Vogt, B. A., & Paxinos, G. (2014). Cytoarchitecture of mouse and rat cingulate cortex with human homologies. *Brain Structure & Function*, *219*(1), 185-192. https://doi.org/10.1007/s00429-012-0493-3

Wang, Z., & Busemeyer, J. R. (2016). Interference effects of categorization on decision making. *Cognition*, *150*, 133-149. https://doi.org/10.1016/j.cognition.2016.01.019

Wiehler, A., Branzoli, F., Adanyeguh, I., Mochel, F., & Pessiglione, M. (2022). A neuro-metabolic account of why daylong cognitive work alters the control of economic decisions. *Current Biology*, *32*(16), 3564-3575.e5. https://doi.org/10.1016/j.cub.2022.07.010

Wyart, V., & Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, *11*, 109-115. https://doi.org/10.1016/j.cobeha.2016.07.003

## APPENDIX

### Bayesian cue denoising

We consider the following situation: An agent has to make a perceptual decision based on an noisy cue. In our experimental setting, this cue corresponds to a set of 10 lines that are either vertical or horizontal (e.g. 3 horizontal and 7 vertical lines). The correct response depends on the direction of the majority of the line (e.g. left for a majority of horizontal lines and right for a majority of vertical lines). Importantly, we manipulate decision difficulty through the difference between the number of vertical and horizontal lines ($\Delta L$ , hereafter).

We suppose that as the agent samples the cue, it receives a noisy, Gaussian-distributed, evidence signal $S_t$ centered on $\Delta L$ (see Figure 2):

$$p(S_t|\Delta L) = N(S_t; \Delta L, \sigma_L^2)$$

where $\sigma_L^2$ is the variance of the noise.

We also suppose that the agent starts with a prior belief about $\Delta L$ :

$$p(\Delta L) \; = \; N(\Delta L; \mu_0, \sigma_0^2)$$

Without loss of generality, we set the agent's prior mean to $\mu_0 = 0$.
Bayes' rule then yields the posterior belief about $\Delta L$ after having observed the sequence of noisy evidence $S_{1:t}$ :

$$p(\Delta L|S_{1:t}) = N(\Delta L, \mu_t, \sigma_t^2)$$
$$= \; N(\Delta L; \mu_0, \sigma_0^2) \prod_{\tau=1}^{t} N(S_\tau; \Delta L, \sigma_L^2)$$

where the mean and variance of the posterior estimate of $\Delta L$ at time $t$ are given by:

$$\sigma_t^2 = \left( \frac{1}{\sigma_0^2} + \frac{t}{\sigma_L^2} \right)^{-1}$$

$$\mu_t = \left( \frac{\sigma_L^2}{t \times \sigma_0^2} + 1 \right)^{-1} \times \frac{1}{t} \sum_{\tau=1}^{t} S_\tau$$

Note that, as cue sampling time increases, $\mu_t$ converges to the sample mean of the noisy signal, because $\lim_{t \to \infty} \frac{\sigma_L^2}{t \times \sigma_0^2} = 0$. This guarantees that decisions based upon its posterior belief exhibit expected error rates that monotonically decrease with cue sampling time.

**Agnostic Evidence Accumulation Model**

The Bayesian denoising model above can be used to inform the agent's metacognitive predictions about future changes in its belief state. However, we rather suppose that such metacognitive evaluation is agnostic about the way cue evidence is integrated. Instead, it simply assumes that spending more time on controlled processing of cue information (i) decreases the posterior variance, and (ii) increases the expected absolute change of the posterior mode. These assumptions can be simply written as follows:

$$\sigma_t^2 = \left(\frac{1}{\sigma_0^2} + \beta t\right)^{-1}$$
$$\mu_t = \mu_0 + \delta_t,\ \delta_t \sim N(0, \gamma t)$$

where efficacy parameters $\beta$ and $\gamma$ can be to the previously described Bayesian cue denoising scenario:

$$\beta = \frac{1}{\sigma_L^2}$$

$$\gamma \simeq \frac{1}{T_{\max}} \sum_{t=1}^{T} \frac{\sigma_L^2 + t\sigma_0^2}{\left(\frac{\sigma_L^2}{\sigma_0^2} + t\right)^2}$$

In the absence of default preferences, we assume that the agent answers on the side that corresponds to its mean estimate. This simplifies the derivation of the agent's confidence, i.e. the subjective probability of responding correctly:

$$P(\text{correct}|\mu_t, \sigma_t^2) = P(\text{sign}(\Delta L) = \text{sign}(\mu_t)|\mu_t, \sigma_t^2))$$
$$= \begin{cases} P(\Delta L \geq 0|\mu_t, \sigma_t^2) \text{ if } \mu_t \geq 0 \\ P(\Delta L < 0|\mu_t, \sigma_t^2) \text{ otherwise.} \end{cases}$$

which can be approximated as follows:

$$P(\text{correct}|\mu_t, \sigma_t^2) = s\left(\frac{\pi|\mu_t|}{\sqrt{3}\sigma_t}\right)$$

where $s(x) = \frac{1}{1+\exp(-x)}$ is the standard sigmoid mapping.

**Markov Decision Process and ensuing Bellman equation**

At this point, we ask when the agent should stop investing cognitive control and commit to a decision, given the expected costs and benefits of cognitive control. As we will see, this issue is an instance of optimal stopping problem, which we will solve using a discrete-time, finite-horizon Markov Decision Process (MDP).

We define the states of the MDP as follows:

$\forall t \in \{0, \cdots, T\}$, $s_t = (\mu_t, \sigma_t^2, u_t)^\intercal \in \mathbb{R} \times \mathbb{R} \times \{0,1\}$ is the state vector. Here, $\mu_t$ is the instantaneous evidence (i.e. the above posterior mean about $\Delta L$), $\sigma_t^2$ is the posterior variance $\sigma_t^2$, and $u_t$ is a flag indicating if the stopping action has already been previously chosen at a previous time point.

We define the admissible agent's actions as follows:

$\forall t \in \{0, \cdots, T\}$, $a_t \in \{0,1\}$, where 0 and 1 correspond to stopping (committing to a decision) and continuing cue sampling, respectively.

At each time step, the state-action couple $(s_t, a_t)$ is associated with the following recurrent (stochastic) dynamics:

$$\sigma_{t+1}^2 = \left(\frac{1}{\sigma_0^2} + \beta\right)^{-1}$$
$$\mu_{t+1} = \mu_t + \delta,\ \delta \sim N(0, \gamma)$$
$$u_{t+1} = u_t a_t$$

with initial state $s_0 = (0, \sigma_0^2, 1)^\intercal$.

Without loss of generality, let us assume that the net benefit associated with a state-action couple $(s_t, a_t)$ is only realized at the time $t_{\text{stop}}$ when the agent commits to a decision ($a_{t_{\text{stop}}} = 0$). We call this time step $t_{\text{stop}}$ the stopping time. At this point, we assume that this reward function incorporates a benefit term $B$ that is function of $s_t$ and a cost term polynomial in $t$, as follows:

$$Q_t(s_t, a_t) = [B(s_t) - \alpha_C t^{\nu_C}]\delta_{a_t,0}\delta_{u_t,1}$$

where $u_t = 0$ for $t > t_{\text{stop}}$.

Now, this agent aims at maximizing its expected cumulated net benefit, starting from its current state until it reaches a finite time horizon $T$. Let $V_\tau^*(s_\tau)$ be the optimal value function, i.e. the maximal expected cumulated net benefit if the agent is in state $s_\tau$ at time $\tau$ (where $0 \leq \tau \leq T$ ):

$$V_\tau^*(s_\tau) = \max_{\{a_t\}_{\tau \leq t \leq T}} \left\{ \mathrm{E}\left[\sum_{t=\tau}^{T} Q_t(a_t, s_t)\right]\right\}$$

When $\tau < T$ , we can take the first term ($t = \tau$) out of the sum and the expectation since it doesn't depend on any random variable, and the remaining expectation term becomes now conditional on $s_\tau$ and $a_\tau$:

$$V_\tau^*(s_\tau) = \max_{\{a_t\}_{\tau \le t \le T}} \left\{ \mathrm{E}\left[ Q_\tau(s_\tau, a_\tau) + \sum_{t=\tau+1}^{T} Q_t(a_t, s_t) \right] \right\}$$
$$= \max_{\{a_t\}_{\tau \le t \le T}} \{ Q_\tau(s_\tau, a_\tau) + \mathrm{E}[\textstyle\sum_{t=\tau+1}^{T} Q_t(a_t, s_t) \,|s_\tau, a_\tau] \}$$

where the term within the conditional expectation is simply the optimal value function of $s_{\tau+1}$ conditional on $s_\tau$ and $a_\tau$ . We thus obtain the Bellman equation:

$$V_\tau^*(s_\tau) = \max_{a_\tau} \{ Q_\tau(s_\tau, a_\tau) + \mathrm{E}[V_{\tau+1}^*(s_{\tau+1}) | s_\tau, a_\tau] \}$$

Now, we can turn to the derivation of the optimal policy $\pi_\omega^*$ that maximizes the expected cumulated net benefit of the agent for this MDP by applying a backward induction procedure.

**Derivation of the Optimal Policy**

At time $\tau = T$ , the agent has no other choice than stopping, so by convention:

$$\pi_\omega^*(s_T, T) = 0$$

Here, the optimal value is simply the net benefit from stopping at time $T$ :

$$V_T^*(s_T) = Q_T\big(s_T, \pi_\omega^*(s_T, T)\big) = Q_T(s_T, 0)$$

At time $\tau = T - 1$ , Bellman equation yields the following optimal policy:

$$\pi_\omega^*(s_{T-1}, T-1) = \mathrm{argmax}_{a_{T-1} \in \{0,1\}} \{ Q_{T-1}(s_{T-1}, a_{T-1}) + \mathrm{E}[V_T^*(s_T) | s_{T-1}, a_{T-1}] \}$$
$$= \mathrm{argmax}_{a_{T-1} \in \{0,1\}} \{ Q_{T-1}(s_{T-1}, a_{T-1}) + \mathrm{E}[Q_T(s_T, 0) | s_{T-1}, a_{T-1}] \}$$

Let's now do a case disjunction and treat separately the cases where $a_{T-1} = 0$ and $a_{T-1} = 1$. We get:

$$Q_{T-1}(s_{T-1} a_{T-1}) = \delta_{a_{T-1},0} Q_{T-1}(s_{T-1}, 0) + \delta_{a_{T-1},1} Q_{T-1}(s_{T-1}, 1)$$
$$= \delta_{a_{T-1},0} Q_{T-1}(s_{T-1}, 0) + 0$$

and:

$$\mathrm{E}[Q_T(s_T, 0) | s_{T-1}, a_{T-1}]$$
$$= \delta_{a_{T-1},0}\, \mathrm{E}[Q_T(s_T, 0) | s_{T-1}, a_{T-1} = 0]$$
$$+ \delta_{a_{T-1},1}\, \mathrm{E}[Q_T(s_T, 0) | s_{T-1}, a_{T-1} = 1]]$$
$$= 0 + \delta_{a_{T-1},1}\, \mathrm{E}[Q_T(s_T, 0) | s_{T-1}, a_{T-1} = 1]$$

where the expectation is taken under the MDP state transition density (here only posterior mode dynamics are stochastic: other MDP states are deterministic).
By grouping the corresponding disjunctions together, we get the optimal policy:

$$\pi_\omega^*(s_{T-1},) = \text{argmax}_{a_{T-1}\in\{0,1\}}\{\delta_{a_{T-1},0}(Q_{T-1}(s_{T-1},0)+0), \delta_{a_{T-1},1}(0 + \text{E}[Q_T(s_T,0)|s_{T-1},a_{T-1}=1])\} = 0 \text{ if } Q_{T-1}(s_{T-1},0) \geq \text{E}[Q_T(s_T,0)|s_{T-1},a_{T-1}=1]$$
$$1 \text{ else.}$$

The value function at time $T-1$ is thus given by:

$$V_{T-1}^*(s_T) = \max\{Q_{T-1}(s_{T-1},0), \text{E}[Q_T(s_T,0)|s_{T-1},a_{T-1}=1]\}$$

At time $\tau = T-2$ , by the same reasoning we get the optimal policy:

$$\pi_\omega^*(s_{T-2},T-2) = \text{argmax}_{a_{T-2}\in\{0,1\}}\{\delta_{a_{T-2},0}(Q_{T-2}(s_{T-2},0)+0) + \delta_{a_{T-2},1}(0+\text{E}[V_{T-1}^*(s_{T-1})|s_{T-2},a_{T-2}=1])\} = 0 \text{ if } Q_{T-2}(s_{T-2},0) \geq \text{E}[V_{T-1}^*(s_{T-1})|s_{T-2},a_{T-2}=1]$$
$$1 \text{ else.}$$

Thus, the value function at time step $T-2$ is:

$$V_{T-2}^*(s_{T-2}) = \max\{Q_{T-2}(s_{T-2},0), \text{E}[V_{T-1}^*(s_{T-1})|s_{T-2},a_{T-2}=1]\}$$

More generally, for $\forall t \in \{0,\cdots,T-1\}$, we have the following optimal policy:

$$\pi_\omega^*(s_t,t) = 0 \text{ if } Q_t(s_t,0) \geq \text{E}[V_{t+1}^*(s_{t+1})|s_t,a_t=1]$$
$$1 \text{ else.}$$

And we have the following value function:

$$V_t^*(s_t) = \max\{Q_t(s_t,0), \text{E}[V_{t+1}^*(s_{t+1})|s_t,a_t=1]\}$$

This concludes the derivation of the optimal policy.

**Defining the Benefit term under default policies**

We assume that the agent possesses a default policy that may vary over trials but doesn't depend on the accumulated evidence within a decision trial. A simple possibility for the default

policy is reinforcement learning or win-stay/lose-switch heuristics (which effectively proxies reinforcement learning at high learning rates). Under this default policy, left and right actions are given values $Q_{\text{default}}^{\text{Left}}$ and $Q_{\text{default}}^{\text{Right}}$, respectively.

We also suppose that the agent faces a credit assignment - or attribution - problem, regarding the hidden mechanism that generates reward outcomes. At the very least, two alternative scenarios are possible: either reward follows from selecting a given action (direct reward contingency), or it follows from properly orienting action according to the cue (indirect reward contingency). Within a Bayesian framework, this credit assignment problem can be solved by monitoring both action-outcome and cue-outcome contingencies over trials, and then comparing the ensuing marginal likelihoods. Note that the task design induces ambiguous attribution, essentially because of corrective trials and noisy cue-related evidence processing. We assume that, by the end of the training phase, mice converge on a stable probability estimate for both direct and indirect reward contingencies. In what follows, $\lambda$ denotes the mice's subjective estimate of the probability that actions are directly rewarded (the so-called "attribution belief"). For example, if $\lambda = 0$ , the agent fully attributes her payoff to aligning her actions with the cue. For intermediate values of $\lambda$, the agent has some credence for each of the scenarios.

At a given time point $t$ , the agent is in a state $s_t$ :

If $\lambda = 0$ , then the value of the Left action simply is the reward the agent expect to get from responding Left:

$$Q_{\text{cue}}^{\text{Left}}(s_t) = R \times s\left(\frac{\pi\mu_t}{\sqrt{3}\sigma_t}\right)$$

where $R$ is the reward at stake, which is scaled by the agent's confidence in obtaining the reward from responding Left.

If $\lambda = 1$ , then the value of the Left action is simply $Q_{\text{default}}^{\text{Left}}$.

At the decision time, the benefit $B(s_t)$ is obtained by marginalizing over the agent's attribution belief, and taking the action with the highest value:

$$B(s_t) = \max\left\{(1-\lambda)Rs\left(\frac{\pi\mu_t}{\sqrt{3}\sigma_t}\right) + \lambda Q_{\text{default}}^{\text{Left}}, (1-\lambda)Rs\left(\frac{-\pi\mu_t}{\sqrt{3}\sigma_t}\right) + \lambda Q_{\text{default}}^{\text{Right}}\right\}$$

This yields the following expression for the full net benefit:

$$Q_t(s_t, a_t) = \left[\max\left\{(1-\lambda)Rs\left(\frac{\pi\mu_t}{\sqrt{3}\sigma_t}\right) + \lambda Q_{\text{default}}^{\text{Left}}, (1-\lambda)Rs\left(\frac{-\pi\mu_t}{\sqrt{3}\sigma_t}\right) + \lambda Q_{\text{default}}^{\text{Right}}\right\} - \alpha_C t^{\nu_C}\right]\delta_{a_t,0}\delta_{u_t,1}$$