Tracking subjects' strategies in behavioural choice experiments at trial resolution

Silvia Maggi[1], Rebecca M. Hock[1], Martin O'Neill[1,2], Mark J. Buckley[3], Paula M. Moran[1,4], Tobias Bast[1,4], Musa Sami[5] and Mark D. Humphries[1*]

1. School of Psychology, University of Nottingham, UK.
2. Dept of Health & Nutritional Sciences, Atlantic Technological University, Sligo, Ireland
3. Department of Experimental Psychology, University of Oxford, Oxford, UK.
4. Neuroscience@Nottingham, University of Nottingham, UK.
5. Institute of Mental Health, University of Nottingham, UK

* Contact: mark.humphries@nottingham.ac.uk

Abstract

Investigating the strategies engaged by subjects in decision making and learning requires tracking their choice strategies on a trial-by-trial basis. Here we present a simple but effective probabilistic approach to tracking choice strategies at trial resolution, using Bayesian evidence accumulation. We show this approach identifies both successful learning and the exploratory strategies used in decision tasks performed by humans, non-human primates, rats, and synthetic agents. We find learning occurs earlier and more often than estimated using classical approaches. Also, win-stay and lose-shift strategies, often considered as complementary, are consistently used independently, with the adoption of lose-shift preceding both learning new rules and switching away from old rules. Our approach can be extended to any discrete choice strategy, and its low computational cost is ideally suited for real-time analysis and closed-loop control.

Introduction

Experiments on decision making typically take the form of discrete trials in which a subject is required to make a choice between two or more alternatives (Packard and McGaugh, 1996; Yin and Knowlton, 2004; Behrens et al., 2007; Churchland et al., 2008; Buckley et al., 2009; Hanks and Summerfield, 2017; Izquierdo et al., 2017; Juavinett et al., 2017). The subject's goal is to learn by trial and error the target rule leading to the correct choice (Mansouri et al., 2020), guided by feedback alone (reward and/or error signals; Ito et al. (2015); Amarante et al. (2017)) or with additional predictive stimuli (such as texture or odour; van Wingerden et al. (2010); Wang et al. (2019); Banerjee et al. (2020)). Such experimental designs can be equally applied whether the subject is primate (Rudebeck et al., 2008; Buckley et al., 2009; Leeson et al., 2009; Shiner et al., 2015), rodent (Raposo

et al., 2012; Brady and Floresco, 2015; Tait et al., 2018; Campagner et al., 2019; Harris et al., 2021), or insect (Giurfa and Sandoz, 2012).

Neuroscience is increasingly looking to fine-grained analyses of decision-making behaviour across species (Krakauer et al., 2017; Pereira et al., 2020), seeking to characterise not just the variation between individuals, but also an individual's variability across trials (Smith et al., 2004; Roy et al., 2021; Ashwood et al., 2022), in order to better match neural activity to behaviour. Thus, when analysing the behaviour of each subject in a choice task, we ideally want to know both when that subject has learnt the correct rule, and precisely what strategies the subject used. To find the answers to these questions, for any trial we want to be able to ask: what choice strategy are they using now?

Classic approaches to analysing performance on decision-making tasks, like psychometric curves, are computed post-hoc and implicitly assume a fixed strategy throughout (Kim and Shadlen, 1999; Carandini and Churchland, 2013). Choice behaviour is though inherently non-stationary: learning, by definition, requires changes to behaviour, humans and other animals switch strategies as they explore tasks, and experimenters switch target rules, introducing new rules or switching previous reward contingencies (Birrell and Brown, 2000; Genovesio et al., 2005; Rich and Shapiro, 2007; Leeson et al., 2009; Donoso et al., 2014; Brady and Floresco, 2015; Jang et al., 2015; Powell and Redish, 2016; Bartolo and Averbeck, 2020; Russo et al., 2021). Some algorithms can detect changes in choice behaviour that correspond to learning (Smith et al., 2004; Suzuki and Brown, 2005), but they lack information on what strategies subjects use to learn or whether the subject's learnt strategy matches the target rule.

We introduce a simple but effective Bayesian approach to inferring the probability of different choice strategies at trial resolution. This can be used both to infer when learning has occurred, by tracking the probability of the strategy matching the target rule, and to infer which exploratory strategy or strategies the subject used during learning. We show that our inference algorithm successfully detects strategies used by a synthetic agent, then show how our approach can infer both successful learning and exploratory strategies across decision-making tasks performed by rats, non-human primates, and humans. The analysis provides general insights about learning: that win-stay and lose-shift, commonly discussed as though yoked together (e.g. Izquierdo et al., 2017; Miller et al., 2017; International Brain Laboratory et al., 2021; Roy et al., 2021), are dissociable strategies; and that subjects consistently adopt lose-shift immediately before learning a new rule or stopping the execution of a rule that is no longer rewarded. Our approach is computationally efficient, could be used in real-time for triggering changes to task parameters or neural stimulation, and is easily extended to more complex decision-making tasks: to aid this we provide a publicly available toolbox in MATLAB (`https://github.com/Humphries-Lab/Bayesian_Strategy_Analysis_MATLAB`) and Python (`https://github.com/Humphries-Lab/Bayesian_Strategy_Analysis_Python`).

Results

Given a subject performing a decision-making task with two or more choices, our goal is to compute the probability $P(\text{strategy}_i(t)|\text{choices}(1:t))$ that a particular strategy i has been used on current trial t, given the subject's history of choices up to and including trial t.

With this probability, we ask two questions about choice strategies:

1. What have the subjects learnt? By defining strategy i as one of the available rules in the task, we can test if the strategy the subject is using aligns with the rewarded rule,

with high values for $P(\text{strategy}_i(t)|\text{choices}(1:t))$ indicating successful acquisition of the rule. This approach can be equally applied to the early learning of a task during training and to the behaviour of well-trained subjects to track their responses to rule switching (for example, to see which other rules they switch between).

2. What strategies are being tried in order to learn the rule? Subjects can try a range of strategies in order to explore the structure of the environment. For example, animals often show some combination of win-stay or lose-shift behaviour at the start of learning (Jang et al., 2015; Akrami et al., 2018; Constantinople et al., 2019). By choosing strategy i to be one of these exploratory strategies, we can track its occurrence. Potential exploratory strategies include random guessing, history-dependence, and information-seeking.

The number n and type of strategies are selected by the experimenter, and comparing the probabilities of the strategies ($P(\text{strategy}_1(t)) \dots P(\text{strategy}_n(t))$) determines the strength of support for each strategy on trial t.

Computing strategy probabilities at trial resolution

We compute $P(\text{strategy}_i(t)|\text{choices}(1:t))$ using Bayes theorem to provide an estimate of each probability and the uncertainty of that estimate, given the available evidence up to trial t:

$$\underbrace{P(\text{strategy}_i(t)|\text{choices}(1:t))}_{\text{posterior}} \propto \underbrace{P(\text{choices}(1:t)|\text{strategy}_i(t))}_{\text{likelihood}} \times \underbrace{P(\text{strategy}_i(t))}_{\text{prior}}, \tag{1}$$

where the posterior is the estimate of the probability of strategy i being executed, which we want to obtain; the likelihood is the consistency of the choices with the strategy we are assessing; and the prior is the initial estimate of the probability of using that strategy. We now define each of the likelihood, prior, and posterior distributions.

On each trial there are two outcomes for a given strategy i: either the choice made is consistent with that strategy (a success, $x = 1$) or it is not (a failure, $x = 0$). The history of choices [choices$(1:t)$] is then the sequence of successes or failures to execute the strategy. As each trial has two outcomes the likelihood $P(\text{choices}(1:t)|\text{strategy}_i(t))$ is a binomial distribution.

Setting the prior as the Beta distribution $B(\alpha_i, \beta_i)$, defined in the range $[0, 1]$, allows us to treat the unknown probability of strategy i as a random variable. The Beta distribution is the so-called conjugate prior for the binomial distribution, meaning that the posterior is also a Beta distribution, dramatically simplifying the calculations in Equation 1.

Indeed, so simple that updating the posterior distribution $P(\text{strategy}_i(t)|\text{choices}(1:t))$ on trial t becomes, in practice, elementary arithmetic (Methods): given the parameters of the Beta distribution $(\alpha_i(t-1), \beta_i(t-1))$ for strategy i on the previous trial, and the choice on trial t is either consistent ($x = 1$) or not ($x = 0$) with strategy i, then update $\alpha_i(t) \leftarrow \alpha_i(t-1)+x$ and $\beta_i(t) \leftarrow \beta_i(t-1)+(1-x)$. The initial prior is defined by the values set for $\alpha(0)$ and $\beta(0)$. This, then, provides a full trial-by-trial estimate of the posterior probability of strategy i, in the Beta distribution $B(\alpha_i(t), \beta_i(t))$, using the accumulated evidence up to trial t.

Using Bayes theorem in this way assumes that the parameter being estimated – here $P(\text{strategy}_i(t)|\text{choices}(1:t))$ – is stationary. But choice behaviour in a decision-making task is likely to be non-stationary, as when the subject switches exploratory strategies while learning, or when the experimenter changes the target rule after learning, forcing

the subject to switch away from the learnt strategy. Naive use of the Bayesian approach will fail to track these changes.

The solution we adopt here is to weight the evidence entered into the Bayesian update by its recency. We keep a running total of past successes to execute strategy i up to trial t, exponentially weighted by how far in the past each success occurred: $s_i^*(t) = \sum_{\tau=1}^{t} \gamma^{t-\tau} x(\tau)$, where $\gamma \in (0, 1]$ is the rate of evidence decay. Similarly, we keep a running total of past failures to execute strategy i: $f_i^*(t) = \sum_{\tau=1}^{t} \gamma^{t-\tau}(1 - x(\tau))$. We then obtain the following iterative update rules for trial t (Methods):

- Observe whether the choice on trial t was consistent ($x(t) = 1$) or not ($x(t) = 0$) with the execution of strategy i;
- Update the running totals of evidence, decaying the prior evidence: $s_i^*(t) = \gamma s_i^*(t - 1) + x(t)$; and $f_i^*(t) = \gamma f_i^*(t - 1) + (1 - x(t))$;
- Update the parameters for the posterior distribution: $\alpha_i(t) \leftarrow \alpha_i(0) + s_i^*(t)$ and $\beta_i(t) \leftarrow \beta_i(0) + f_i^*(t)$.

With $\gamma = 1$ we recover standard Bayesian inference; the lower γ, the stronger recent history is weighted. Throughout we use $\gamma = 0.9$ unless otherwise noted; we show below that strategy inference is quite robust to the choice of γ, and we consider this point further in the Discussion.

Figure 1 demonstrates the calculation and interpretation of strategy probability using example data from a naive rat performing consecutive sessions on a Y-maze task. In every trial, the rat had to choose either the left or right arm (Figure 1a), and received reward based on four rules that were applied in sequence (right arm, cued arm, left arm, uncued arm), with rules switching after 10 consecutive correct trials or 11 correct out of 12 trials. Figure 1b plots the time-course of the relevant task and behavioural variables needed to infer the rat's strategy: its choices, rewards, and the cue location. We used these to compute $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the strategies corresponding to the first three rules: 'go right', 'go cued', and 'go left'. Figure 1c plots the resulting posterior distributions for those three strategies across three sequential trials during the learning of the 'go left' rule. These show how the rat consistently choosing the left arm updates (α, β) for $P(\text{go left})$ to shift it rightward, while the rat choosing not to go down the right arm when it is cued updates (α, β) for $P(\text{go cued})$ to shift it leftward; as a consequence, 'go left' becomes the most probable strategy.

We use here two summaries of the full posterior probability distribution $B(\alpha(t), \beta(t))$ to explore changes over trials. First, for each trial, we can take the best estimate of the probability that strategy i is being executed as the *maximum a posteriori* (MAP) value of $B(\alpha(t), \beta(t))$ (Figure 1c). Plotting the MAP probabilities for each trial in Figure 1d shows how they capture three key dynamics of choice behaviour: first, when rules are learnt, as the probability of the target rule's strategy becomes large and dominant; second, when learnt strategies persist after the target rule has been changed – for example, 'go right' persists as the dominant strategy for over a hundred trials after the reinforced rule has switched to 'go cued' around trial 25; and, third, when subjects switch strategies, as when the rat switches from 'go right' to 'go cued' in the session beginning at trial 152, emphasising that choice strategy is non-stationary. A Bayesian approach without the decay of evidence cannot track rule learning or strategy switching in these example data (SI Figure S1). With the decay of evidence, the estimated MAP probabilities, and hence inferred changes in strategy, are robust to the choice of prior distribution (SI Figure S2).

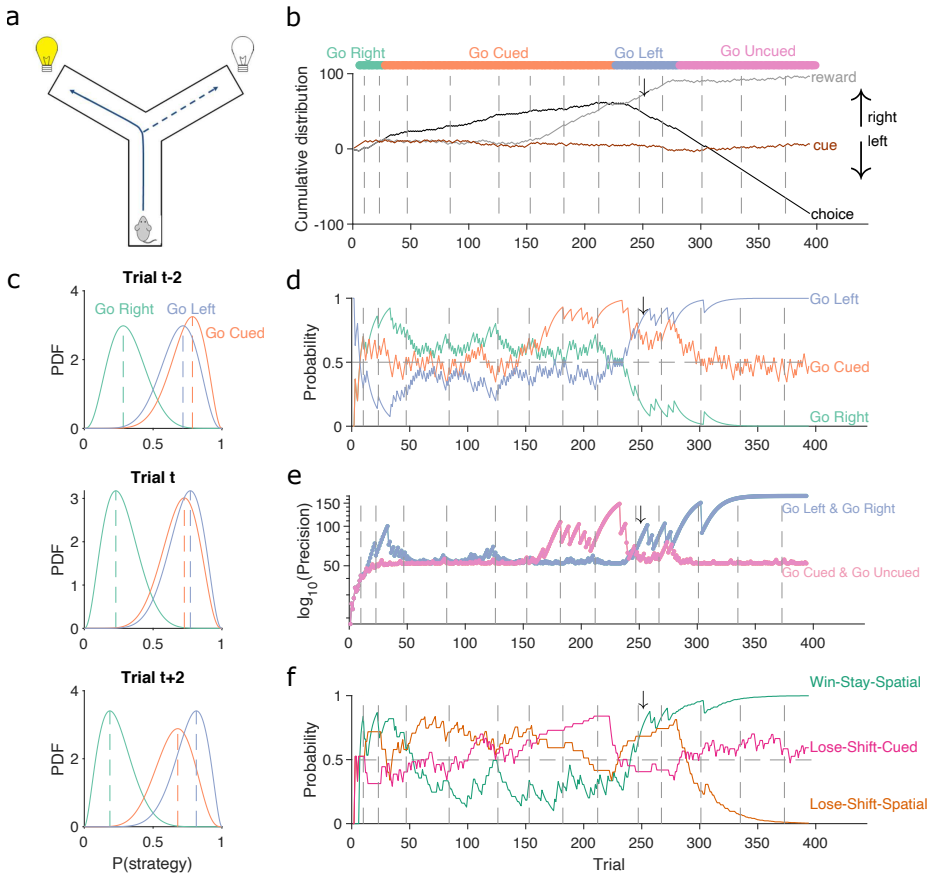

Figure 1: A Bayesian approach to detecting strategies.
(**a**) Schematic of the Y-maze task. The rat receives a reward at the arm end if it chose the correct arm according to the currently enforced rule. A light cue is randomly switched on at the end of one of the two arms on each trial. Rules were in the sequence 'go right', 'go cue', 'go left', 'go uncued', and switched after ten consecutive correct trials or 11 out of 12.
(**b**) Example rat performance on the Y-maze task. We plot performance across 14 consecutive sessions: vertical grey dashed lines indicate session separation, while coloured bars at the top give the target rule in each trial. Performance is quantified by the cumulative distributions of obtained reward (grey) and arm choice (black). We also plot the cumulative cue location (brown), to show it is randomised effectively. Choice and cue distributions increase by +1 for right and decrease by -1 for left. Small black arrow indicates the trial t shown in panel c.
(**c**) Example posterior distributions of three rule strategies for three trials. Trial t is trial 250 in panel b. Vertical dashed coloured lines identify the *maximum a posteriori* (MAP) probability estimate for each strategy.
(**d**) Time-series of MAP probabilities on each trial for the three main rule strategies, for the same subject in panel b. Horizontal grey dashed line indicates chance. We omit 'go uncued' for clarity: as it is the complementary strategy of 'go cue' (P(go uncued) = 1-P(go cued)), so it is below chance for almost all trials.
(**e**) Precision (1/variance) for each trial and each of the tested strategies in panel d; note the precisions of mutually exclusive strategies (e.g. go left and go right) are identical by definition.
(**f**) Time-series of MAP probabilities for three exploratory strategies. Staying or shifting is defined with respect to the choice of arm ('Spatial') or the state of the light in the chosen arm ('Cued') – for example if the rat initially chose the unlit arm and was unrewarded, then chose the lit arm on the next trial, this would be a successful occurrence of 'Lose-Shift-Cued'.

Second, we can also summarise the amount of evidence we have for our MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ by computing the precision – the inverse of the variance – of the full probability distribution: the higher the precision, the more concentrated the Beta distribution is around the MAP estimate of the probability of strategy i (Figure 1e).

The above focused on strategies corresponding to the target rules to learn. We can equally define models for strategies that are explorative, to capture the strategies the subject engages while learning the rule. For example, we may want to assess a simple win-stay strategy. A model for win-stay based on choice would be: count a success if trial $t-1$ was correct and the subject made the same choice on trial t; count a failure if trial $t-1$ was correct and the subject made a different choice on trial t. In Figure 1f, we plot MAP probabilities for three such explorative strategies, win-stay and lose-shift for the prior choice of left or right arm, and lose-shift based on the prior cue position. This demonstrates when exploratory strategies are used to guide behaviour; for example, of these strategies, lose-shift in response to the choice of arm begins to dominate from around trial 50 to 150 (see Figure 1d).

Throughout we distinguish rule strategies, which correspond to one of the reinforced rules available in the task, from exploratory strategies, which are all other possible choice strategies given task information. The choice of strategies is down to the user (Methods).

Robust detection of strategies in synthetic data

To demonstrate that our approach can accurately infer known strategies, and the effect of evidence decay γ, we simulated an agent working on a two-alternative forced choice task. The rewarded rule was to choose the randomly cued option (brown curve in Figure 2a). The agent switched its strategy every 100 trials, across five strategies, with only the fourth strategy corresponding to the rewarded rule (Figure 2a). This set of implanted strategies included a range of options examined in behavioural studies, often individually, and we used many consecutive trials with the same strategy to explore the full range of γ. We inferred $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for eight strategies, including the five implanted ones, using the agent's choice data.

Our inference algorithm accurately tracked all strategy switches, giving the correct strategy the highest MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ immediately after the agent switches to it (Figure 2b; SI Figure S3 shows full MAP probability estimates for all strategies and values of γ). Figure 2b also illustrates that, after the agent switched its strategy at 100 and 300 trials, our estimate of the dominant strategy from the MAP probability alone became ambiguous, as two strategies converged on $P(\text{strategy}_i(t)|\text{choices}(1:t)) = 1$. This reflects that the evidence available to us, the observer, was itself ambiguous between these two strategies. Nonetheless, from the MAP probability estimate alone, our algorithm clearly identified the alternatives, and gives strong evidence of which strategies were not being used. Moreover, this ambiguity could be resolved by examining the evidence for each strategy.

The confidence in our estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ was not identical between the two possible strategies after trials 100 and 300 (Figure 2c). This suggests a simple decision rule for selecting the most likely strategy being executed at trial t: choose strategy i with the highest MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$, resolving ties by choosing the strategy with the maximum precision (Figure 2d). Such a decision rule collapses a rich set of information into a binary choice, useful for visualisation and simple analysis.

Nonetheless, it performs well (Figure 2e), correctly identifying the strategy in more than 88% of trials across a wide range of choices for the rate of evidence decay γ (grey

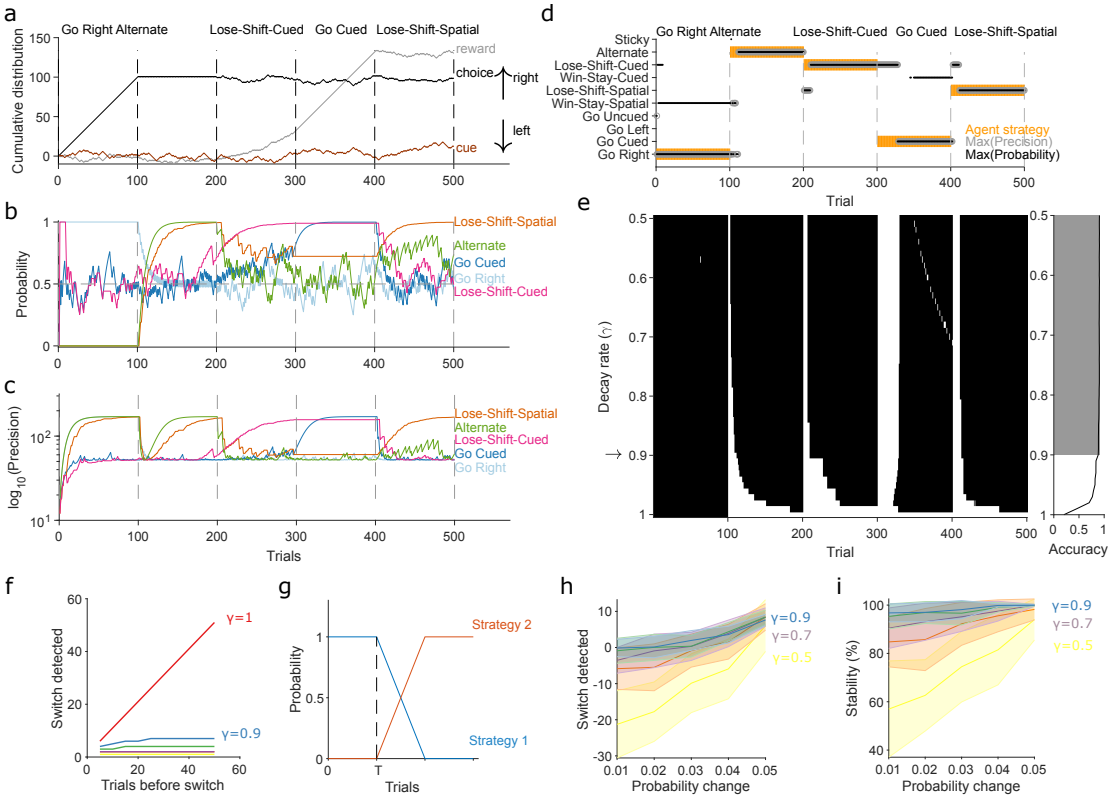

Figure 2: Robust inference of strategies in synthetic data.
(**a**) Cumulative distribution of raw behavioural data for the synthetic agent, performing a two-alternative task of choosing the randomly-cued option. Curves are as in Figure 1a. Vertical dashed lines indicate strategy switches by the agent.
(**b**) MAP values of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the five implanted strategies across trials, for $\gamma = 0.9$. Chance is 0.5.
(**c**) Precision for the five implanted strategies.
(**d**) A decision rule. For every tested strategy, listed on the left, we plot for each trial the strategies with the maximum MAP probability (black dots) and maximum precision (grey circles) across all strategies, and the agent's actual strategy (orange bars). Maximising both the MAP probability and precision uniquely identifies one tested strategy, which matches the agent's strategy.
(**e**) Robustness of the decision rule across evidence decay rates. In the left-hand panel, we plot whether the agent's strategy was successfully inferred (black) or not (white) by the decision rule. The right-hand panel shows the accuracy in detecting the correct strategy for each γ; grey shading shows γ values within the top 5% for accuracy. Most switches were detected in ten trials or fewer across this range of γ.
(**f**) Number of trials until a strategy switch is detected, as a function of the number of trials using the first strategy. One line per decay rate (γ).
(**g**) Schematic of a gradual switch between two strategies after trial T. Trial T is 500 in panels h and i, so that past-history effects have plateaued, giving worst-case performance for detection. Detection of the switch is measured from when the probabilities cross over: when the probability of executing strategy 2 exceeds that of executing strategy 1.
(**h**) Number of trials after the crossover trial until detection, as a function of how quickly the probabilities of executing each strategy change per trial. Negative trials indicate detection before the crossover of probabilities. Plots show means and SD over 50 repeats; $\gamma = 1$ omitted as detection scales with T when there is no evidence decay.
(**i**) Stability of the detected switch: the proportion of trials after the detection for which the algorithm reported the new strategy (strategy 2) was dominant. Plots show means and SD over 50 repeats.

shading in Figure 2e). Moreover, performance decays gracefully as γ increases further. Figure 2e also illustrates nicely why the decay of evidence is essential, as with $\gamma \approx 1$, switches in strategy could not be detected; but with $\gamma \approx 0.9$ or lower, most switches were detected in ten trials or fewer.

To show these insights generalised, we conducted extensive simulations of how the algorithm performed when detecting a switch from one arbitrary strategy to another. We first examined abrupt switches, like those used in the example synthetic task (Figure 2a-e), where a first strategy is used consistently for T trials, then a new strategy is used consistently thereafter. We deemed a switch detected on the first trial for which the algorithm's MAP probability estimate of the new strategy was greater than the MAP probability estimate of the first strategy. If we did not decay evidence over time it took another T trials to detect the switch (Figure 2f, $\gamma = 1$); but with evidence decay abrupt switches were detected orders-of-magnitude faster, and at a rate that rapidly plateaued with increasing T (Figure 2f). Decaying evidence thus ensures that there is little impact of how long a first strategy was used on the algorithm's ability to detect a switch to a new strategy. And while stronger evidence decay (smaller γ) increased the speed at which the abrupt switch was detected, even for $\gamma = 0.9$ the worst-case performance was ten trials (Figure 2f).

We then examined how well the algorithm detected gradual switches between strategies, modelled by linearly changing the probabilities of executing the first and new strategies between 0 and 1 after trial T (Figure 2g); we measured how quickly a switch was detected from when the two probabilities of execution crossed over. Smaller γ apparently led to faster detection of the switch on average, which scaled with how gradually the probabilities of the two strategies changed – the slower the change, the faster the detection (Figure 2h). Yet these faster detections occurred before the probabilities of using each of the two strategies had crossed over (negative values in Figure 2h), suggesting that the detection with smaller γ was due to a chance run of trials with the new strategy. To examine this, we looked at the trials after the switch to the new strategy was first detected and asked how stably the new strategy was detected thereafter. Smaller γ led to less stable identification of the new strategy, whereas $\gamma = 0.9$ was highly stable (Figure 2i). Our use of $\gamma = 0.9$ here is thus a compromise between the speed and stability of updating strategy probabilities.

We also examined abrupt and gradual switches to and between conditionally deterministic strategies, i.e. strategies whose execution depends on a task condition being met, such as win-stay. These switches could take two forms. First, detecting a switch between strategies that have the same condition (e.g. "win") depends on how frequently that condition is met. Defining the probability of the condition being met on trial t as $p(\text{condition met}(t))$ (e.g. $p(\text{win}(t))$), then the above results on the number of trials needed to detect abrupt and gradual switches would all scale by $1/p(\text{condition met}(t))$. Second, when a new condition occurs then a new conditional strategy can be adopted and be detected. Our simulations showed that while the number of trials needed to detect a new conditional strategy must scale with $p(\text{condition met}(t))$, detection is typically within five trials if the condition occurs on at least half of all trials and can be in as little as one trial (SI Figure S4). Our Bayesian algorithm can thus rapidly detect switches in rule or exploratory strategies under a wide range of forms that a switch could take.

There is a potential issue that tested exploratory strategies do not include the true strategy or strategies used by a subject. We thus sought to place bounds on how we expect the inference algorithm to behave if the true strategies are missing. We defined $p(\text{match})$ as the proportion of trials on which a tested strategy would be expected to match the choice

made by the subject's true strategy; we could then interpret the MAP probability of a tested exploratory strategy as an estimate of $p(\text{match})$ (SI Figure S5). Enumerating the possible values of $p(\text{match})$ for a task with n choices revealed three bounds (SI Figure S6; see Methods for details of enumeration). First, that if the algorithm estimates the MAP probability of all tested exploratory strategies to be $\leq 1 - 1/n$ then the true strategy or strategies are missing. Second, that a MAP probability $> 1 - 1/n$ for a tested exploratory strategy is evidence for at least a partial match to the true strategy (for example, in Figure 2b, when the true strategy adopted by the agent was lose-shift-spatial the tested strategy of alternate had a MAP probability of around 0.75, because both strategies will give a change of choice after a loss, thus matching on all loss trials, but not all rewarded trials). Indeed, one way of interpreting exploratory strategies is that they provide evidence about what task features, such as obtaining reward (or not), and the subject's response to them, such as staying or shifting, are driving exploration; the algorithm's estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ is thus a read-out of that evidence. Third, that a MAP probability approaching 1 is strong evidence that the tested strategy is, or is equivalent to, the true strategy.

Collectively these results show that obtaining trial resolution probabilities gives us rich information about the potential choice strategies on each trial. From this we can unambiguously identify one strategy from many, and otherwise strongly identify strategies that are not being used, even if we use a simple decision rule that collapses full posterior distributions to a binary outcome. With these insights to hand, we now apply our inference algorithm to uncover the choice strategies used in tasks performed by rats, humans, and monkeys.

Trial resolution inference reveals evidence of earlier learning

Classic approaches to detecting learning typically depend on arbitrary criteria based on trial counts, whether seeking a sequence of correct trials or percentage correct over a window of trials (e.g. Birrell and Brown, 2000; Boulougouris et al., 2007; Floresco et al., 2008; Leeson et al., 2009; Peyrache et al., 2009; Brady and Floresco, 2015). Such criteria are also typically conservative, requiring long sequences or high percentages to give unambiguous evidence of learning. We thus asked whether our trial-resolution inference could provide a more principled and more accurate approach to detecting learning, by focusing on just the trial-by-trial changes in probability for the strategy corresponding to the target rule.

To investigate this question, we used data from rats learning cross-modal rule shift tasks. We compared data from two conceptually similar tasks, one involving the choice of one of two arms in a Y-maze (Figure 1a, data from Peyrache et al. 2009) and the other the choice of one of two levers in an operant box (Figure 3d; task design from Brady and Floresco 2015). In both tasks, the rat had to learn at least one spatial rule based on the position of the choice option (arm or lever) and at least one cue-driven rule, based on the position of a randomly-illuminated cue light. As noted above, the Y-maze task contained a sequence of up to four rules, two spatial (left/right) and two cue-driven. In the data available to us, three rats experienced all four rules and one rat experienced the first two rules, giving a total of seven shifts from spatial to cue-driven rules and three shifts from cue-driven to spatial rules. In the lever-press task, rats were trained on one spatial and one cue-driven rule: 16 rats experienced a shift from a spatial to a cue-driven rule, and 12 experienced the shift from a cue-driven to a spatial rule (Methods). We thus computed $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the strategies corresponding to the target rules in each task.

Classic definitions of learning the target rule were originally used in both tasks. In the

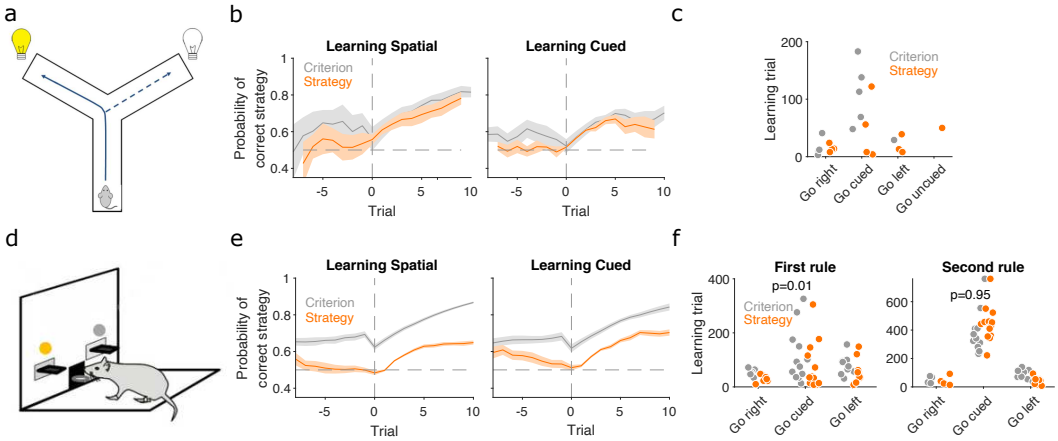

Figure 3: Detecting learning
(**a**) Schematic of Y-maze task, from Figure 1a.
(**b**) MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the correct strategy for the spatial (left panel) and cued (right panel) rules on the Y-maze, aligned to the learning trial for each subject (trial 0, vertical dashed line). Learning trials are defined by the experimenters' original criterion (grey) or by our strategy criterion (orange). Curves represent means (bold line) $\pm$ SEM (shading) for N sessions meeting each criterion across the 4 rats (Spatial rule: original criterion $N = 5$, strategy criterion $N = 7$; Cued rules: original criterion $N = 5$, strategy criterion $N = 5$). Horizontal dashed line is chance.
(**c**) Comparison between learning trials (dots) identified based on the original criterion (grey) or on the strategy criterion (orange) for each Y-maze rule. Trial zero is the first trial with that rule enforced. Rules are labelled in the order they were presented to the rats.
(**d**) Schematic of the lever-press task. Rats experienced both spatial (e.g. choose left lever) and cued (e.g. choose lit lever) rules.
(**e**) As for panel b, for the lever-press task. $N = 28$ rats.
(**f**) Similar to panel c, comparing the learning trials identified based on the original criterion or our strategy criterion, for the first (left) or second (right) rule learnt by each rat in the lever-press tasks. Each dot is one rat. P-values are from a Wilcoxon signed rank test between the learning trials identified by the original and by the strategy criterion ($N = 28$ rats). For the first rule, the learning trials based on the original and strategy criterion are 88.9 ± 12.3 and 59 ± 11.3 mean$\pm$SEM, respectively. For the second rule, the learning trials based on the original and strategy criterion are 224.3 ± 33 and 272 ± 33 mean$\pm$SEM, respectively.

Y-maze study, learning was defined post-hoc by three consecutive correct trials followed by 80% correct until the end of the session, with the first of the three trials then labelled as the learning trial (Peyrache et al., 2009). In the lever-press task, learning was defined on-line as ten consecutive correct trials, after which the session ended (Brady and Floresco, 2015) – the first of these ten we label as the learning trial. However, when we computed the probability of the strategy corresponding to the target rule, we found it was often above chance before these classically-defined learning trials for both the Y-maze (Figure 3b, 'Criterion') and lever-press tasks (Figure 3e, 'Criterion'), suggesting rules were learnt notably earlier than detected.

We thus assessed an alternative strategy-based criterion to identify the learning trial: the first trial at which the probability for the target-rule's strategy remained above chance (0.5) until the end of the session (Methods). Using this criterion, the change in probability around the learning trial showed clearer steps from chance to above-chance performance in both tasks (Figure 3b,e, 'Strategy').

In the Y-maze task, this strategy-based criterion identified that learning occurred 12

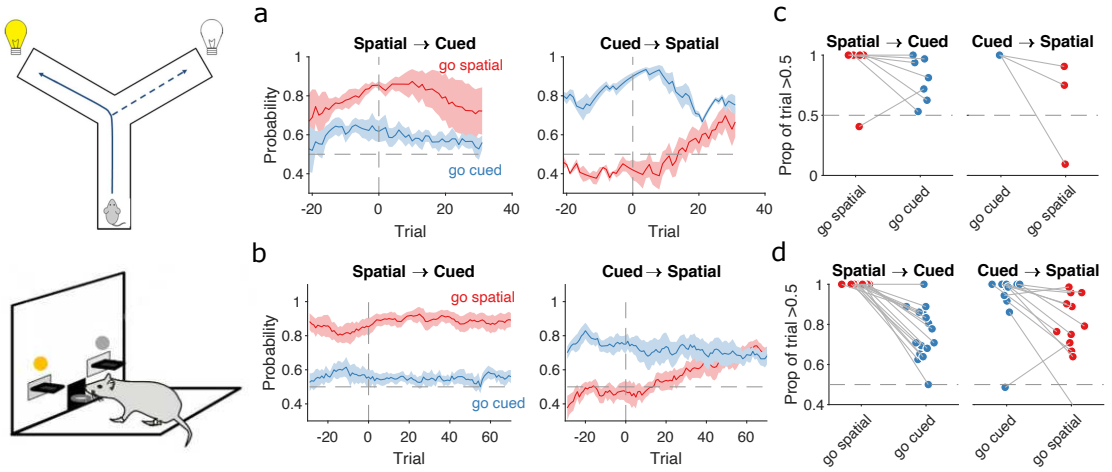

Figure 4: Detecting responses to task changes
(**a**) Responses to rule switches in the Y-maze task. The left panel plots the MAP estimates of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the target spatial and cue strategies aligned to the switch from a spatial to a cued rule (N=7 sessions, from 4 rats). The right panel plots the MAP estimates aligned to the switch from a cued to a spatial rule (N=3 sessions, from 3 rats). For both panels, curves plot means ± SEM (shading) across sessions, aligned to the rule switch trial (vertical dashed line). The horizontal dashed line indicates chance.
(**b**) Same as panel a, but for the lever-press task. $N = 16$ rats for spatial→cued; $N = 12$ cued→spatial.
(**c**) Response to each rule-switch in the Y-maze task. Each dot shows the proportion of trials after a rule-switch in which the labelled strategy was above chance. Proportions calculated for a window of 30 trials after the rule switch. Lines join data-points from the same switch occurrence.
(**d**) Same as panel c, for the lever-press task. Proportions calculated for a window of 60 trials after the rule switch, to take advantage of the longer sessions in this task.

times across the four rats, compared to 10 times using the original criterion (Figure 3c). In particular, unlike the original criterion, the strategy criterion suggested that all three rats which experienced the 'go left' rule learnt it, and one rat learnt the 'go uncued' rule.

For the lever-press task, we could compare the learning trials identified by both approaches because every rat was trained to the original criterion. The strategy criterion robustly identified learning earlier, on average 30 ± 5.8 trials earlier, during learning of the first target rule (Figure 3f, left). By contrast, both criteria identified similar learning trials during the second target rule each rat experienced (Figure 3f, right). They thus both agreed that learning the cued rule after a spatial rule took longer than learning a spatial rule after the cued rule (Figure 3f, right).

Flexibility in responses to rule changes depends on rule type

Learning a target rule is one of two key events in decision-making tasks. The other is responding to a change in the target rule. Changes in behaviour around rule switches are useful to understand the flexibility of behaviour, and the relative dominance of one mode of responding (e.g. spatial) over another (e.g. cue-driven) (Birrell and Brown, 2000; Floresco et al., 2008; Buckley et al., 2009; Wang et al., 2019). We thus used our inference algorithm to examine trial resolution changes in choice strategy around rule-switches in the same Y-maze and lever-press tasks, to show how our approach can quantify behavioural flexibility.

In both tasks, rats which had been training on a spatial rule continued to have a high

probability of executing a spatial strategy on each trial long after the rule had switched to a cue-based target rule (Figure 4a,b left panels). Thus, rats showed little behavioural response during switches from spatial to cued rules.

The switch from a cued to a spatial rule in both tasks evoked a considerably faster change in strategy. Over the same time window as the spatial→cued switch, while the probability of selecting the cued strategy declined slowly, the mean probability of selecting the now-correct spatial strategy increased above chance within 20 trials of the rule switch (Figure 4a,b right panels).

To demonstrate the fine-grained analysis possible with our Bayesian algorithm, we also quantified the variability in rule-switch responses across rats by calculating for each rat the proportion of trials in which each rule strategy was above chance after the switch (Figure 4c,d). The sustained use of the now-incorrect strategy after both spatial→cued and cued→spatial rule switches could be seen for individual rats (Figure 4c,d). In both tasks, the reduction in use of the 'go cued' strategy after cued→spatial rule switches was also evidenced by that strategy reverting to chance for a few trials (that is, the proportion of trials above chance was less than one: Figure 4d, right). Rats in both tasks also showed considerable variation in how often the 'go spatial' strategy was above chance after the switch to the spatial rule (Figure 4c,d, right), suggesting widely-varying rates of adapting to the rule switch.

Our trial resolution algorithm has thus provided evidence from both learning (Fig 3f) and responding to rule switches (Fig 4) that the spatial→cued rule switch is harder for rats to detect than the cued→spatial switch, across two different tasks. Moreover, the behavioural response to the cued→spatial rule switch (Figure 4a,b, right panels) shows that, as both were above chance, the newly correct strategy ('go spatial') can be acquired even while the now incorrect strategy ('go cued') is still actively used.

Learning and rule-shifts are driven by lose-stay, not win-shift

Having used the changes in rule strategies to characterise both learning and the response to rule shifts, we then sought to understand the exploratory strategies that were driving these changes.

To probe how learning was achieved, we computed $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for four exploratory strategies based on the task features pertinent to the target rules: for the spatial rules, win-stay and lose-shift based on the choice of arm or lever; for the cued rules, win-stay and lose-shift based on the cue position. We also looked at outcome-insensitive strategies of repeating or of alternating the previous choice. In both the Y-maze and lever-press tasks, rule learning was preceded by a higher probability of lose-shift in response to the corresponding feature of the target rule (choice or cue), with win-stay to that feature only growing in probability after the identified learning trial (Figure 5a,d; for all exploratory strategies see SI Figure S7). Consistent across both rule types and both tasks, this result suggests that lose-shift and win-stay are distinct exploratory choice strategies, with lose-shift driving learning.

To further understand what drove the asymmetries between the responses to spatial→cued and cued→spatial rule switches (Figure 4a-b), we examined $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the above exploratory strategies around these rule switches (Figure 5b-c, e-f; results for all exploratory strategies are plotted in SI Figure S8). In both tasks the high probability of continuing to enact the spatial strategy after the spatial→cued switch could be attributed to the continuing dominance of win-stay to the choice of arm (Figure 5b, left) or lever (Figure 5e, left), even though this feature was no longer relevant to the task.

Yet the exploratory strategies showed evidence of immediate latent responses to the

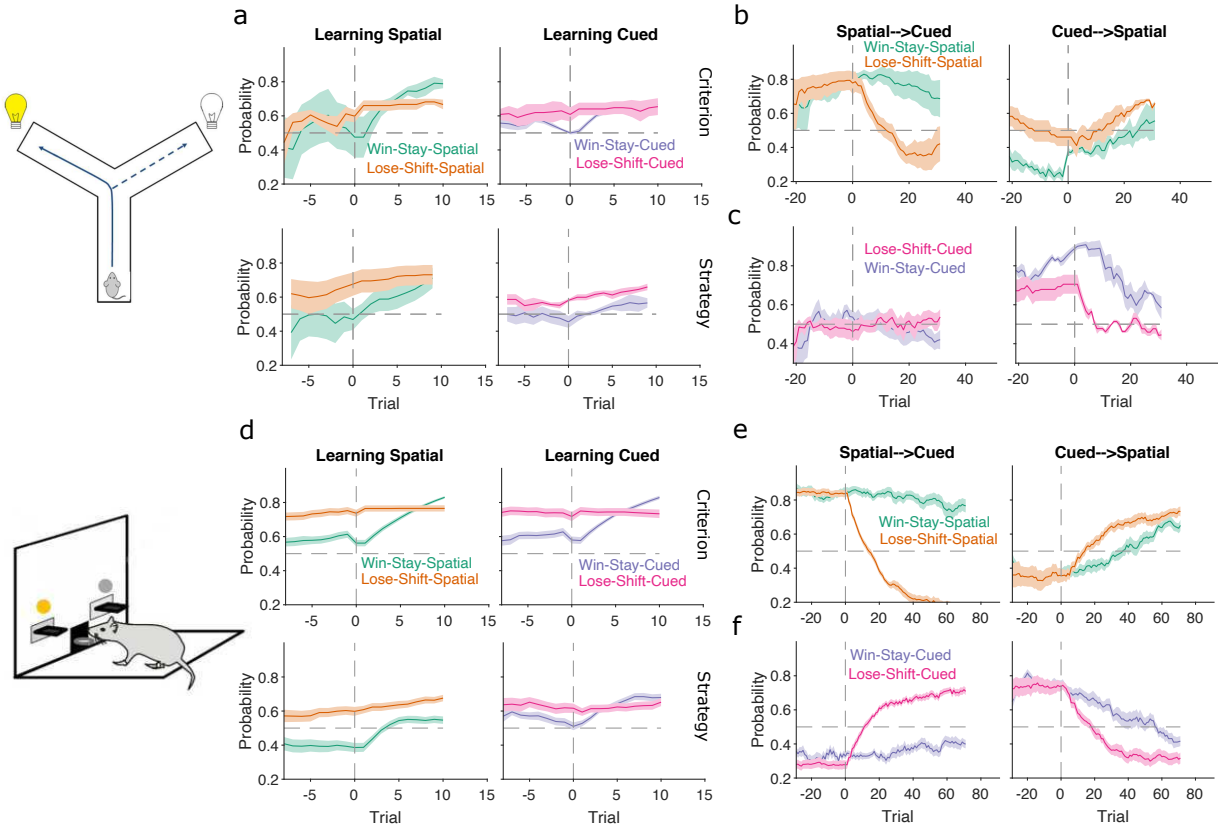

Figure 5: Lose-shift drives learning and responses to rule changes
(**a**) Changes in the probability of win-stay and lose-shift strategies around learning. We plot probabilities for choice-driven (left) and cue-driven (right) forms of win-stay and lose-shift aligned to the identified learning trial (trial 0). Learning trials identified by either the original study's criterion (top) or our rule-strategy criterion (bottom). Lines and shading show means ± SEM (N given in Figure 3b).
(**b**) Rule-switch aligned changes in probability for choice-based win-stay and lose-shift strategies (N given in Figure 4a).
(**c**) Rule-switch aligned changes in probability for cue-based win-stay and lose-shift strategies.
(**d**-f) Same as panels a-c, but for the lever-press task. Panel d: $N = 28$ rats per curve. Panel e-f: $N = 16$ rats spatial→cued; $N = 12$ cued→spatial.

spatial→cued rule switch: in both tasks, lose-shift to the choice of arm (Figure 5b, left) or lever (Figure 5e, left) fell below chance rapidly after the rule switch, suggesting a latent awareness of the irrelevance of the spatial feature. And in the lever-press task, within about 20 trials the rats had a high probability of lose-shift behaviour in response to the cue (Figure 5f, left), so showing some latent learning of part of the cued rule; but without the corresponding win-stay to the cue, which would have been necessary to acquire the full rule.

Similarly, the asymmetric changes to the cued and spatial strategy choices around the cued→spatial rule switch (seen in Figure 4) could be attributed to independent changes in lose-shift and win-stay strategies. The rapid increase in 'go spatial' was driven by lose-shift to the choice of arm or lever rapidly increasing above chance after the switch (Figure 5b,e, right). By contrast, the slower dropping of 'go cued' was due to the probability of win-stay in response to the cue remaining above chance after the switch (Figure 5c,f, right). Yet the probability of lose-shift in response to the cue rapidly fell to chance (Y-maze: Figure 5c, right) or below it (lever-press: Figure 5f, right) after the switch, further evidence of some

latent awareness that the rule had switched. Collectively, these results emphasise that win-stay and lose-shift are independent strategies, and illustrate how our trial-resolution approach can uncover latent learning.

Latent exploratory strategies disguise the absence of learning in humans

We now turn to demonstrating insights from the inference algorithm on a wider range of tasks. Here we analyse human performance on a simple stimulus-choice task, with no rule switches, to illustrate how the algorithm performs with only a few trials; in the next section we consider non-human primate performance on a stimulus-action task with stochastic reward.

Human participants performed a two-choice task, similar to that of Pessiglione et al. (2006), in which they had to choose between two stimuli that appeared above and below a central fixation point (Figure 6a), across three pairs of stimuli corresponding to three interleaved trial types. The top or bottom position of each pair of stimuli was randomised on each trial. In 'gain' trials, one stimulus increased total reward with a probability of 0.8, the other nothing; in 'loss' trials, one stimulus reduced total reward with a probability of 0.8, the other nothing; and in 'look' trials, participants randomly chose one of the two stimuli, with no associated outcome. Their implicit goal was thus to maximise reward by learning to select the gain stimulus and avoid the loss stimulus. As participants only experienced 30 of each trial type, these data could present strong challenges to classical approaches of identifying if and when participants learnt to correctly choose the gain stimulus or avoid the loss stimulus, and what strategies they adopted during learning. By contrast, our Bayesian approach can identify strategies within a few trials (Figure 2).

To identify learning, we separately computed $P(\text{strategy})$ for the strategies of selecting the gain stimulus and of avoiding the loss stimulus (Figure 6b). Despite the few trials involved, these estimates were not sensitive to the choice of prior (Figure 6b and SI Figure S9). Across all participants, the MAP estimates of $P(\text{strategy})$ showed evident learning of choosing the gain stimulus, but not of avoiding the loss stimulus (Figure 6c). Applying the same learning criterion as above, of MAP probability being above 0.5 until the end of the session, our inference algorithm suggests most participants learnt to choose the gain stimulus (13/20), but not to avoid the loss stimulus (5/20) (Figure 6d). Even then, the few who did putatively learn to avoid the loss stimulus did so only towards the end of the 30 trials (Figure 6e). Our trial resolution probabilities have thus shown both strong evidence of which participants learnt the gain trials, and that there is weak evidence at best of any participant learning to avoid the loss stimulus.

To further examine this issue, we asked if the apparent learning to avoid loss could be explained by an exploratory strategy. For the participants who putatively learnt either the gain or loss rule strategies we computed $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for a range of exploratory strategies that captured either responding to previous choices (stimuli) or to the previous locations of the chosen stimulus. Stimulus-driven win-stay and lose-shift were both used for learning gain trials, but not for loss trials (Figure 6f). By contrast, only location-driven exploration was consistently, albeit weakly, above chance for loss trials (Figure 6f). Because this suggested participants could be sensitive to the location of the stimuli, we then also examined location-dependent strategies in the 'look' trials, and found participants were strongly biased towards choosing one location over the other; however, this bias did not correlate with the participants' use of the avoid-loss strategy (SI Figure S9). Together, these results suggest that no participant learnt avoid-loss, and those who appeared to were instead executing a location-driven exploratory strategy that happened to coincide with avoid-loss for the final few trials.

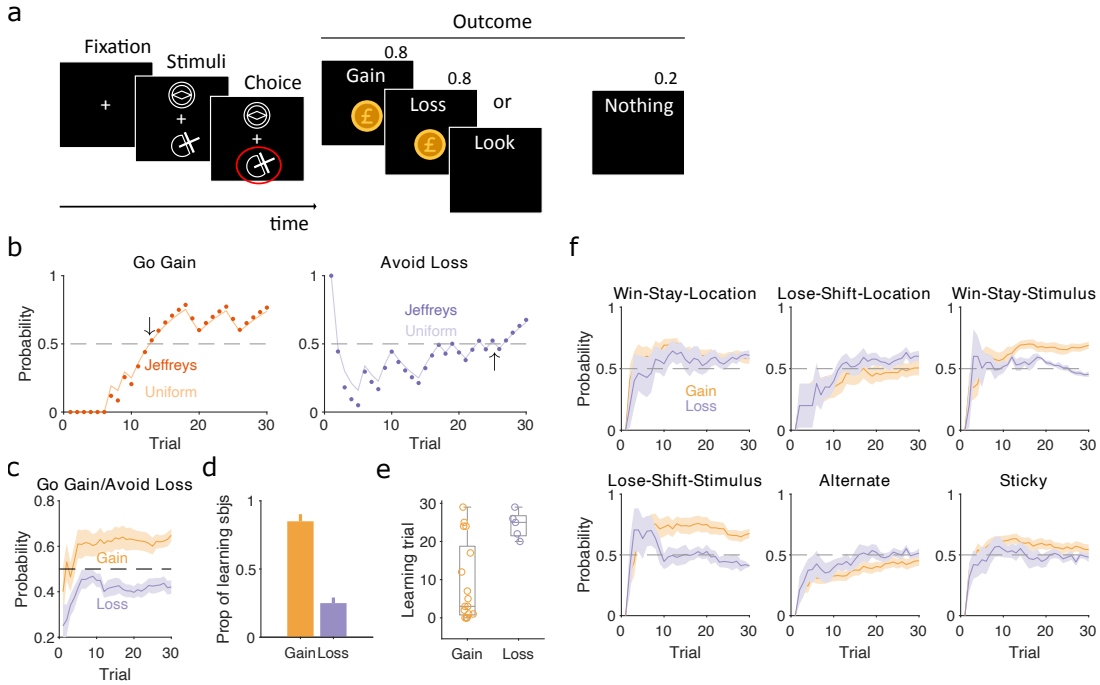

Figure 6: Human learning and exploration on a gain/loss task
(**a**) Gain/loss task. Gain, loss, and look trials each used a different pair of stimuli and were interleaved. A pair of stimuli were randomly assigned to the top or bottom position on each trial. Numbers above Outcome panels refer to the probabilities of those outcomes in the gain and loss trial types.
(**b**) For an example participant, the MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the target rule strategy during gain (left panel) and loss (right panel) trials. We plot two estimates, initialized with our default uniform prior (solid) or Jeffrey's prior (dotted). Horizontal dashed lines indicate chance level. Black arrows indicate the learning trial based on our strategy criterion.
(**c**) The MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ for the target rule strategies across all participants. Curves plot mean $\pm$ SEM (shading) across 20 participants. The dashed line is chance.
(**d**) Proportion of participants that learnt the target strategy for gain or loss trials. Learning was identified by the first trial at which the MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ was consistently above chance until the end of the task (black arrows in panel b). Vertical bar indicates the 95% Clopper-Pearson confidence intervals for binomial estimates.
(**e**) Distributions of learning trials for gain and loss trials, for the participants meeting the learning criterion. Each dot is a subject. Boxplots show the median, and the 25th and 75th percentiles. The whiskers extend to the extreme values.
(**f**) Probabilities of exploratory strategies for the learning participants (N=13 for gain, N=5 for loss). Curves plot mean $\pm$ SEM (shaded area) MAP probabilities.

Choice strategy driven by reward probability not magnitude in a stochastic decision task

More complex decision-making tasks use probabilistic and variable-size rewards for all choices. These present further challenges for the analysis of choice behaviour, not least because the variation of both probability and size of reward increases the number of features upon which subjects may base their choice strategies. To examine these challenges, we used our inference algorithm to analyse a session of data from a rhesus macaque performing a stimulus-primed two-choice task (Figure 7a).

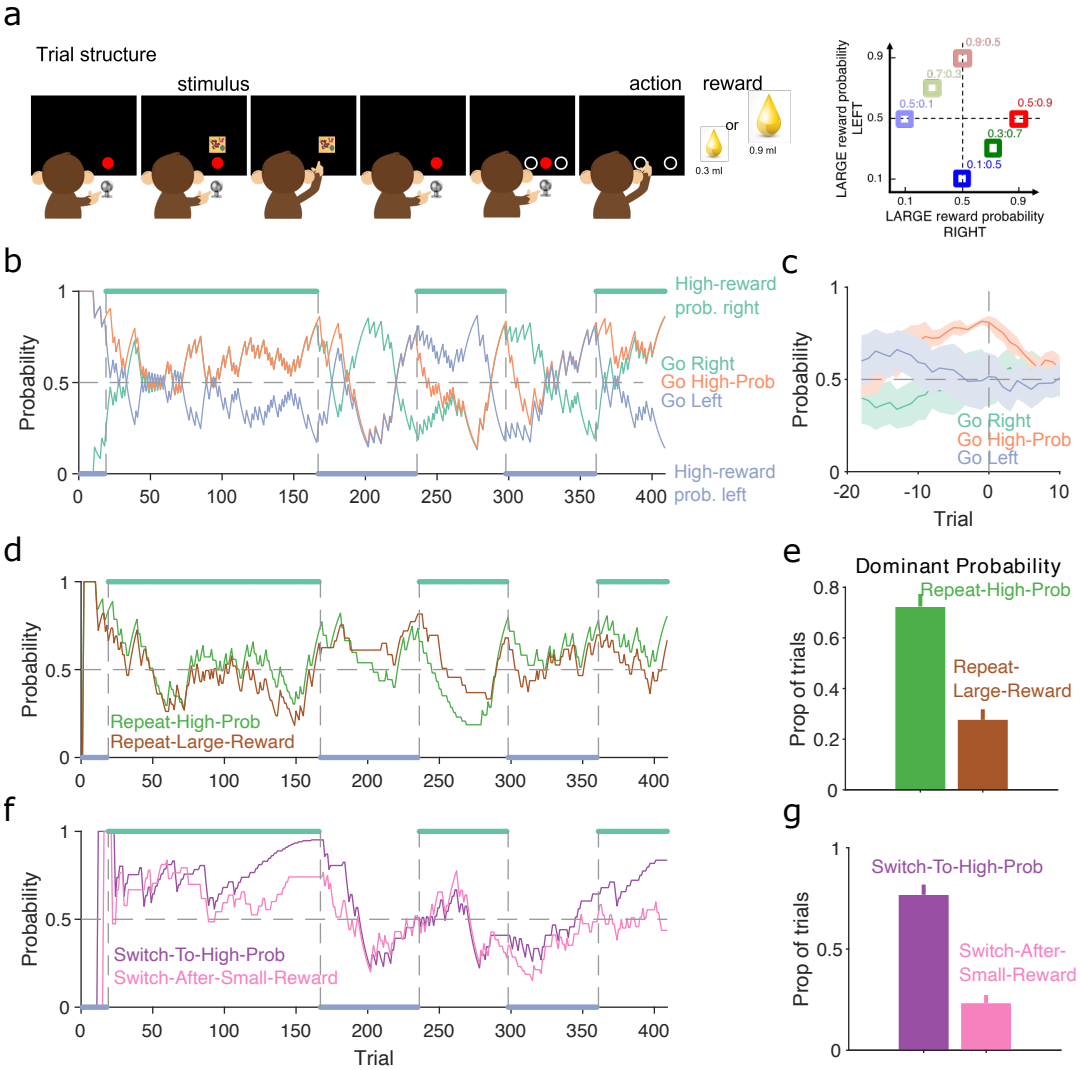

Figure 7: Probabilistic reward

(**a**) Schematic of the stimulus-to-action reward task. Left: The monkey initiated a trial by touching a key after the red dot appeared on-screen. A stimulus appeared, predicting which side of the upcoming choice would have the higher probability of obtaining the large reward. Touching the stimulus and then the key brought up the choice options on-screen. The monkey indicated its decision by touching the appropriate circle on-screen. Right: probabilities of obtaining the large reward for each of the six stimuli.

(**b**) MAP probabilities for 'go right', 'go left' and 'go high-probability side' strategies. Green and blue dots at the top and bottom indicate the side with the higher probability of receiving the large reward; probabilities switched (vertical dashed lines) after the subject chose the high probability option on 80% of trials in a moving window of 20 trials.

(**c**) MAP profiles in panel b aligned to the transition between probability blocks (trial 0). Average (bold lines; n=5 blocks) ± SEM (shaded areas).

(**d**) MAP probabilities along the session for exploratory strategies based on repeating the previous choice after either receiving the large reward or choosing the option with the higher probability of the large reward.

(**e**) Proportion of trials in which the two repeating-choice strategies had the largest MAP probability. Vertical bars are 95% Clopper-Pearson confidence intervals.

(**f**) As panel d, for exploratory strategies based on switching from the previous choice after either receiving the low reward or choosing the option with the lower probability of high reward.

(**g**) As panel e, for the switching strategies in panel f.

In this task, the monkey was presented with a stimulus that predicted the probability of obtaining the larger of two rewards if it subsequently chose either the left or right option on a touchscreen (Figure 7a, left). Six stimuli were associated with a different pair of probabilities for obtaining the larger reward for the left or right choice (Figure 7a, right); whatever the monkey chose, the small reward was delivered if the probability check for the large reward failed. The stimulus and associated pair probabilities switched once the monkey chose the highest-probability option in 80% of trials. With our trial-resolved approach we can immediately ask two questions: did this criterion give sufficient evidence for learning, and did learning that strategy follow the probability of obtaining the large reward or something else?

If the monkey's choices were driven by the highest probability of the large reward alone (or, equivalently, by the option with the highest expected value), then the choice strategy would be to choose the option with that probability. Computing $P(\text{strategy})$ for the corresponding target rule strategies of choosing the left or right option (Figure 7b) confirmed this criterion showed good evidence of learning, as $P(\text{strategy})$ for the highest-probability option became dominant at least 10 trials before the switch of reward probabilities (orange line Figure 7c). Our analysis also suggests that in at least one block (trials 25 to 160 in Figure 7b), the monkey had learnt to choose the appropriate option long before the block switch.

As the choice of either option in this task could result in the large reward, the monkey may have been basing its exploration not on the probability of the large reward, but simply on the value of the obtained reward. To examine this, we computed $P(\text{strategy})$ for two sets of exploratory strategies: one set looking at whether repeating the same option choice was driven by reward or the probability of reward on the previous trial (Figure 7d); and the other looking at whether switching the option choice was driven by reward or the probability of reward on the previous trial (Figure 7f). The dominance of switching or repeating exploratory strategies evolved over the blocks (Figure 7d,f). Independent of this evolution, both repeating and switching strategies were dominated by the probability of reward rather than the reward obtained (Figure 7e,g), evidence that the monkey was basing its decisions more on reward probability than immediate value. Because each session had new arbitrary stimuli for the six pairs of probabilities, the monkey was here re-learning the mapping between stimuli and probabilities, and so was not trivially choosing from an already-learnt association. Rather, these results suggest that, through its prior training, the monkey had learnt the structure of the probabilities, and so was able to use that abstract knowledge to guide its exploratory choices.

Discussion

Subjects may adopt a range of choice strategies during decision-making tasks, which change during training as they learn a target rule, and as the rule is changed. Uncovering how subjects learn to make decisions is essential for understanding what information drives behaviour and linking neural activity to behaviour. Classical approaches to characterising decision making, such as trials-to-criterion measures (Birrell and Brown, 2000; Peyrache et al., 2009; Brady and Floresco, 2015), psychometric functions (Britten et al., 1992; Palmer et al., 2005; Churchland et al., 2008) or reward curves (Gallistel et al., 2004), cannot track changes over time; algorithms for detecting learning points (Smith et al., 2004; Suzuki and Brown, 2005) and behavioural change-points (Durstewitz et al., 2010; Jang et al., 2015) are useful, but do not show what strategies were used during learning or around changes. To address these issues, we have introduced an approach to

decision-making tasks that fully characterises the probability of choice strategies at trial resolution. Our inference algorithm allows tracking both learning, by defining strategies matching target rules, and how subjects learn, by defining exploratory strategies that may or may not depend on features of the task. We have applied our algorithm to a variety of two-choice tasks, across a range of species, including Y-maze and lever-press tasks in rats, a stimulus-outcome mapping task in humans and a stochastic stimulus-action task in non-human primates.

Across the tasks we analysed, our algorithm showed that lose-shift and win-stay are independent exploratory strategies, because a change in the probability of one being executed is not mirrored by a change in the probability of the other. This challenges views that win-stay and lose-shift are complementary strategies, which are often yoked together (Miller et al., 2017; Izquierdo et al., 2017; International Brain Laboratory et al., 2021; Roy et al., 2021). Emphasising this independence, our data show that the high probability of adopting a lose-shift strategy precedes both successful learning and abandoning a previously successful rule, both in Y-maze and lever-press tasks (Figure 5). Intriguingly, this analysis of exploratory strategies also showed evidence of latent learning in both tasks, as after switches from a spatial to a cued rule or vice-versa, rats rapidly stopped using lose-shift in response to choices matching the old rule long before they learnt the new rule.

In both the Y-maze and lever-press tasks the algorithm showed learning occurred earlier and more often than detected in traditional approaches based on counting successful trials in some fixed window. It also showed that, for both tasks, shifting from a spatial to a cue-based rule is harder than the reverse shift, and this difficulty corresponded to the rats showing sustained win-stay in response to being rewarded for their spatial choice of arm in the Y-maze or lever in the lever-press task. Prior work on the same lever press task had shown some evidence of the opposite, that re-learning is faster after switching from a spatial to a cued rule than switching from a cued to a spatial rule (Floresco et al., 2008). Fruitful insights into decision-making could be obtained by future work studying this discrepancy.

There has been considerable recent advocacy for studying the variety of behaviour within and between individuals (Krakauer et al., 2017; Honegger and de Bivort, 2018; Pereira et al., 2020; Roy et al., 2021; Ashwood et al., 2022), both to more richly characterise the diversity of behavioural solutions adopted by humans or other animals facing a task, and as a basis for understanding the link between neural activity and behaviour. As it provides trial-resolution probabilities for each subject, our algorithm also offers a solution for quantifying the behavioural diversity in a given task. We illustrated that potential here by defining learning points for each subject in the Y-maze (Figure 3c), lever-press (Figure 3f), and human gain/loss tasks (Figure 6d,e), and by showing the distribution of exploratory behaviours around rule switches (Figure 4c,d).

Reflecting this urgent need for fine-grained behavioural insights into how subjects make decisions, recent work by Pillow and colleagues introduced two innovative solutions to the problem of tracking a subject's behaviour (Roy et al., 2021; Ashwood et al., 2022). One solution (PsyTrack: Roy et al., 2021) continuously infers the parameters of a psychometric function to track the relative influence of different task features on choice in two-alternative forced-choice tasks; the other (Ashwood et al., 2022) infers the hidden states underlying decision-making behaviour and the transitions between them, by modelling psychometric functions in each state. Their elegant work applies to well-learnt behaviour, using computationally-intensive models fit to all available data post-hoc, but lacks the ability to infer when learning occurs. By contrast, our approach focuses on behavioural strategies adopted during learning and rule switches, accumulates evidence up to trial t so can be

used on-line, and is able to infer strategies in a few tens of trials (Figures 2 and 6), but lacks a model of the transitions between strategies or the parameters which influence the current strategy. An interesting future avenue for our work would be to use the changing probability of different strategies to infer the transition probabilities between them. Our approach and that of Pillow and colleagues are thus complementary, collectively providing a richer toolbox for understanding a subject's behaviour on choice tasks.

The probabilities computed by our algorithm are a description of the observer: what we can tell about the behaviour on trial t from the evidence so far. As we demonstrated, this evidence can sometimes be ambiguous (Figure 2), which is reflected in the high probabilities of two or more strategies. This Bayesian analysis of behaviour does not imply the animal is behaving in any optimal, Bayesian way, nor that its brain is computing or using probabilities (Knill and Pouget, 2004; Fiser et al., 2010; Pouget et al., 2013; Sohn et al., 2019). Nonetheless, using a Bayesian strategy analysis could let us then formulate those questions of optimality in behaviour and look for neural correlates of probabilistic representations. More generally, such a trial-by-trial analysis of strategy could then let us look for the neural correlates of the identified strategies, more accurately tracking the changes in neural activity that precede or align with changes in behavioural strategy (Rich and Shapiro, 2009; Durstewitz et al., 2010; Karlsson et al., 2012; Powell and Redish, 2016; Guise and Shapiro, 2017).

One issue we faced is that Bayesian inference assumes the parameter being estimated is stationary, but choice strategies are not, so any estimate of their probability must also change. Here we solve that problem by weighting the evidence entering the Bayesian update by its recency. This allows tracking the probabilities of strategies as they change, and is robust to the exact weighting used (Figure 2).

The choice of how much to decay the influence of past evidence can determine how quickly a change in strategy is detected (Figure 2), suggesting other solutions to this non-stationary problem are worth pursuing. For example, having a principled way of varying the weighting γ of evidence over time could be useful, so that evidence is rapidly decayed around a switch in strategies and less rapidly during stable execution of a strategy. This would require defining γ as a function of some observable parameter of behaviour. It could also potentially find different evidence decay rates γ between different strategies.

Introducing γ may seem to create a free parameter for the analysis. However, all frequentist approaches to analysing the strategies of subjects also have at least one free parameter, namely the length of the time window over which to compute the probability they are using. This is true both for computing the probability of a particular strategy being used as the proportion of trials it occurs on (Singh et al., 2019; Harris et al., 2021; Trepka et al., 2021) and for classic learning criterion of subjects performing sequential, or a proportion of, successful trials in some time window (e.g. Birrell and Brown, 2000; Boulougouris et al., 2007; Floresco et al., 2008; Leeson et al., 2009; Peyrache et al., 2009; Brady and Floresco, 2015). Sometimes this time window is explicitly chosen (Singh et al., 2019); sometimes the choice is implicit, as the probabilities are computed over trial blocks (Trepka et al., 2021) or phases of the experiment (Harris et al., 2021). Either way, rarely is the influence of that time window considered. Thus, rather than γ introducing a further free parameter, it instead makes the inevitable free parameter of time explicit and bounded: we have placed strong bounds on the values of γ by showing the trade-off between the speed of detecting a new strategy and the stability of detecting it.

Our algorithm provides rich information about choice behaviour by computing full posterior distributions for however many strategies are adopted, for every trial. How that information is then used to inform our understanding of choice behaviour offers a

wealth of development. We here extracted from those distributions a point estimate of the probability of each strategy and the precision of that estimate; other summaries are possible. Further, we suggested here two approaches to interpret the information in the distributions: a decision rule to identify the dominant strategy, and a learning criterion based on the consistent above-chance use of a strategy matching the target rule. These worked well in practice but, as shown with our analysis of humans unable to learn to avoid losses (Figure 6), developing a more nuanced approach to learning criterion could give deeper insights.

The richness of the information depends in part on the choice of strategies examined, which is down to the end-user and what questions they wish to ask of their data. We here distinguished rule strategies and exploratory strategies: while the former should be examined to understand learning, the latter can include a wider range of options than we examined here. Here we focused predominantly on variants of win-stay and lose-shift strategies as these are widely studied (Miller et al., 2017; Constantinople et al., 2019; International Brain Laboratory et al., 2021; Roy et al., 2021; Ashwood et al., 2022), contrasting the different features of a task – subject's spatial choice or the task cues – that define the stay or shift response. But the range of strategy models usable with our inference algorithm is broad enough to cover many potential questions about decision-making: all the algorithm needs from the model is the success or failure to execute that strategy on trial t. As such, our approach is highly extensible. It can be used with more complex strategies than those considered here, such as N-back history strategies (Lau and Glimcher, 2005; Akrami et al., 2018). Pilot or bias tests could be used to estimate the prior distribution's parameters for each subject. And while all tasks here had two choices, the algorithm can just as easily compute probabilities of choice strategies in tasks with three or more choices. Despite this extensibility, the computational cost per trial is extremely low, as elementary arithmetic alone is needed for computing the update to the posterior distribution. Consequently, the inference algorithm we introduce here can assess the probabilities of an extensive, flexible range of choice strategies, and could do so in real-time, allowing detailed interrogation and potentially closed-loop control of animal and human behaviour during decision-making tasks.

Acknowledgements

This work was supported by the Medical Research Council [grant numbers MR/J008648/1, MR/P005659/1 and MR/S025944/1] to M.D.H. and [grant number MR/K005480/1] to M.J.B, and the Biotechnology and Biological Sciences Research Council [grant number BB/T00598X/1] to M.J.B. and M.D.H. S.M. was supported by an Anne McLaren Fellowship from the University of Nottingham. M.S. was supported by a Medical Research Council fellowship and a grant from the Dowager Countess Eleanor Peel Trust. Collection of the data on the rat lever-press task was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) Doctoral Training Programme (DTP) at the University of Nottingham [grant number BB/M008770/1, project 1644954], and we are grateful to Stan Floresco for providing MedAssociates programmes to run the lever-press task. We thank Hazem Toutounji for comments on the manuscript, Nathan Lepora for discussions, and Lowri Powell for contributing to the Python port of the toolbox.

Author contributions

Silvia Maggi: Formal Analysis, Software, Writing – Original Draft Preparation, Visualization, Validation, Data Curation.
Rebecca M Hock: Investigation, Writing – Review and Editing.
Martin O'Neill: Investigation, Writing – Review and Editing.
Mark J Buckley: Supervision, Funding Acquisition, Writing – Review and Editing, Investigation.
Paula Moran: Supervision, Writing – Review and Editing.
Tobias Bast: Supervision, Writing – Review and Editing
Musa Sami: Investigation, Writing – Review and Editing
Mark Humphries: Conceptualization, Methodology, Software, Formal Analysis, Writing - Original Draft Preparation, Visualization, Supervision, Project Administration, Funding Acquisition.

Materials and methods

Code availability

Code toolboxes implementing the algorithm and strategy models are available for MATLAB (`https://github.com/Humphries-Lab/Bayesian_Strategy_Analysis_MATLAB` and Python (`https://github.com/Humphries-Lab/Bayesian_Strategy_Analysis_Python`). These toolboxes will be updated with future improvements to the algorithm.

All code used to produce the figures in this paper is available at `https://github.com/Humphries-Lab/Bayesian_strategy_analysis_Paper`. This repository also includes all data in the paper that we are able to share (see Data Availability Statement below).

Bayesian inference of strategy probability

Here we give the full algorithm as used throughout the paper.

1. Choose a set of N strategy models to test.
2. Each $P(\text{strategy}_i(t)|\text{choices}(1:t))$ is parameterised by a Beta distribution, with parameters $\alpha_i(t), \beta_i(t)$, for $i = 1, 2, \ldots, N$.
3. The prior is defined by the choice of $\alpha_i(0), \beta_i(0)$. We use the uniform prior $\alpha_i(0) = \beta_i(0) = 1$ by default; we use the Jeffrey's prior $\alpha_i(0) = \beta_i(0) = 1/2$ where noted.
4. Initialise the running counts of successful (s_i^*) and failed (f_i^*) trials to use strategy i as $s_i^*(0) = f_i^*(0) = 0$.
5. On every trial t:
   (a) label it for strategy i as either a:
   Success trial choice made on trial t is consistent with strategy i: $x_i(t) = 1$;
   Failure trial choice made on trial t is inconsistent with strategy i: $x_i(t) = 0$,
   Null trial strategy i could not be assessed on trial t.
   (b) For success and failure trials, we update the parameters $\alpha_i(t), \beta_i(t)$:
      i. decay the event totals, and add the new event: $s_i^*(t) = \gamma s_i^*(t-1) + x(t)$; and $f_i^*(t) = \gamma f_i^*(t-1) + (1 - x(t))$;

ii. update the Beta distribution parameters: $\alpha_i(t) \leftarrow \alpha_i(0) + s_i^*(t)$ and $\beta_i(t) \leftarrow \beta_i(0) + f_i^*(t)$.

with a decay rate $\gamma \in (0, 1]$.

(c) For null trials, we assign the previous trial's parameter value: $\alpha_i(t) \leftarrow \alpha_i(t-1)$, $\beta_i(t) \leftarrow \beta_i(t-1)$.

6. Output: time-series of $\alpha_i(t)$ and $\beta_i(t)$ for each strategy i, which define a Beta distribution for each trial.

Null trials occur when a strategy depends on conditions that could not be met in that trial. In our current implementation we interpolate the previous values for α, β for a null trial, to define the probability at which that strategy could have occurred. Future extensions to our work could usefully explore alternative solutions to the handling of null trials.

Task descriptions

Data availability statement

The following data used in this paper are freely available in our GitHub repository `https://github.com/Humphries-Lab/strategy-analysis`:

- Y-maze task data: the 50 sessions we used here, extracted from the full dataset available at crcns.org `http://dx.doi.org/10.6080/K0KH0KH5`;
- Lever-press task: all 32 rats used here
- Primate stimulus-to-action task: the session used here

Data from the human gain/loss task are available from the corresponding author upon reasonable request.

Y-maze – rats

Rat behavioural data on the Y-maze task data came from the study of Peyrache et al. (2009). This dataset includes 50 sessions of four male Long-Evans rats performing a cross-modal rule switch task. Rats were trained to self-initiate the trial by running up the central stem of the maze and choosing one of the two arms. The rats had to learn in sequence one of four rules: 'go right','go cued', 'go left', 'go un-cued'. The cue was a light stimulus randomly switched at the end of one of the two arms. The rule was changed within session after performing 10 consecutive trials or 11 out of 12 correct trials. In the original study, learning was defined as the first of three consecutive trials followed by a performance of $\geq 80\%$ until the end of the session. For each trial, we used the choice location (right or left), the cue location (right or left) and the reward (yes or no) as input data for the Bayesian model.

Lever-press – rats

All experimental procedures were conducted in accordance with the requirements of the United Kingdom (UK) Animals (Scientific Procedures) Act 1986, approved by the University of Nottingham's Animal Welfare and Ethical Review Board (AWERB) and run under the authority of Home Office project license 30/3357.

Rat behavioural data for the lever-press task was taken from 32 male Lister Hooded rats (aged 11 weeks at the start of testing and food-restricted throughout behavioural training), which served as control groups in two micro-infusion studies (each $N = 16$) targeting prelimbic medial prefrontal cortex. These rats had infusion guide cannulae implanted in the medial prefrontal cortex and received saline infusions in the medial prefrontal cortex (for surgery and infusion procedures, see Pezze et al. 2014) at various task stages, including before the sessions during which the data presented in this paper were collected. The task was run in standard 2-lever operant boxes with cue lights above the levers and a food well between them (Med Associates, US) based on a previously published training protocol (Brady and Floresco, 2015).

Rats were pre-trained to press the levers for sucrose rewards. The rats then had to learn an initial rule to receive one reward per correct response: either a spatial rule or a visual cue-based rule. The spatial rule required the rat to select the left or right lever, one of which was designated as the correct response and associated with reward. The visual cue-based rule required the rat to select the lever indicated by a cue light, which appeared above one of the two levers and was always associated with reward irrespective of which lever it appeared above. Each trial ended either when a response was made on a lever or when the 10-s time window for responses had elapsed without a response (defined as an omission). After reaching the criterion of 10 consecutive correct trials, the rats had to learn to shift their responses to the second rule: rats which had learnt the spatial rule first were tested on the cue-based rule, and vice versa.

Half the rats ($N = 16$) experienced the spatial rule first: all reached criterion on it and on the subsequent cued rule. Half ($N = 16$) experienced the cued rule first: 12 reached criterion so were switched to the spatial rule, which all 12 reached criterion on as well. Our strategy-based definition of learning (see below) found the same number of learners in each condition, so we have $N = 28$ rats throughout for analysis of learning (Figure 3) and strategy flexibility (Figure 4).

Gain/loss task – humans

The study was approved by Research Ethics Committee (Stanmore London REC 17/LO/0577). Human participants ($n = 20$, 10 male, 10 female) performed a two choice task, similar to that of Pessiglione et al. (2006), in which they had to choose between two arbitrary stimuli that appeared above and below a central fixation point (Figure 6a), across three pairs of stimuli corresponding to three interleaved trial types. The top or bottom position of each pair of stimuli was randomised on each trial. In 'gain' trials, one stimulus increased total reward by £1 with a probability of 0.8, the other nothing; in 'loss' trials, one stimulus reduced total reward by £1 with a probability of 0.8, the other nothing; and in 'look' trials, participants randomly chose one of the two stimuli, with no associated outcome. 30 trials of each type were presented, randomly interleaved. All participants were informed that the amount they received would be dependent upon their performance, but all received the same amount (£20) at the end of the task.

Stimulus-action task – non-human primates

Rhesus macaques were trained to touch a key in front of a touchscreen monitor when a red dot appeared on the screen (Figure 7a). A visual stimulus then appeared above the dot. After a delay the dot disappeared and the animal had to touch the monitor where the stimulus was present. The dot then reappeared instructing the animal to touch the key again. Two white circles then appeared simultaneously to the left and right of the

dot. After a delay the dot disappeared, and the animal chose to select the left or right circle. A large (0.9 ml) or small (0.3 ml) reward of banana smoothie was delivered after every trial. The rewards were probabilistic and asymmetric to either side (Figure 7a, right panel). There were six reward probability pairings associated with six visual stimuli per session. For example, one visual stimulus predicted the large reward with p = 0.5 following a left action and p = 0.1 following a right action (denoted '0.5:0.1' in Figure 7a). If the probability check for the large reward failed, the small reward was delivered instead. Each of the six stimuli were presented in a blocked design with the stimulus change and block transition occurring after 80% choices (16/20 trials) of the side with the highest probability of the large reward. We analysed here a single session from one monkey.

In the full study from which this session was taken, behavioural data were recorded in two young adult male rhesus macaque monkeys (Macaca mulatta, one aged 8 years, weight 10-13 kg; the other aged 9 years, weight 11-15 kg). All animals in the Buckley lab are socially housed (or socially housed for as long as possible if later precluded, for example, by repeated fighting with cage-mates despite multiple regrouping attempts) and all are housed in enriched environments (e.g. swings and ropes and objects, all within large pens with multiple wooden ledges at many levels) with a 12hr light/dark cycle. The NHPs always had ad libitum water access 7 days/week. Most of their daily food ration of wet mash and fruit and nuts and other treats were delivered in the automated testing/lunch-box at the end of each behavioral session (this provided 'jack-pot' motivation for quickly completing successful session performance; supplemented by the trial-by-trial rewards for correct choices in the form of drops of smoothie delivered via a sipping tube) and this was supplemented with fruit and foraging mix in the home enclosure. For the study the animal was prepared for parallel neurophysiological investigation, and so was head-fixated during the task using a standard titanium head-post (and also had recording implants); implantation adopted standard aseptic surgical methods used in the non-human primate laboratory that are not directly relevant for this behavioural study but are described in detail elsewhere (Wu et al., 2021). All animal training and experimental procedures were performed in accordance with the guidelines of the UK Animals (Scientific Procedures) Act of 1986, licensed by the UK Home Office, and approved by Oxford's Committee on Animal Care and Ethical Review.

Strategy models for each task

The strategy models we use for the analysis of the Y-maze and lever-press task data are defined in Tables 1 and 2. In Figures 3-5 we group the analyses of either 'go left' or 'go right' as 'spatial rules', and the analyses of either 'go cued' or 'go uncued' as 'cued' rules.

Analyses of the human gain/loss task used the same set of exploratory strategies, with the replacement of "spatial" by "location", and "cued" by "stimulus". The two rule strategies were defined as: "go gain" – success if the subject chose the potentially rewarded stimulus; "avoid loss" – success if the subject did not choose the potentially reward-decreasing stimulus.

Analyses of the non-human primate stimulus-action task examined three rule strategies, 'go left', 'go right', and 'go high-probability', the latter defined as a success if the subject selected the option with the higher probability of the large reward. Table 3 defines the exploratory strategy models we used for this task.

Table 1: Rule strategy models for rat tasks. Columns give the conditions to define trial t as a success, failure, or null for each strategy corresponding to a target rule. The rule strategies we consider here do not have null conditions.

	Success	Failure	Null
Go left	chose left option	didn't choose left option	n/a
Go right	chose the right-hand option	didn't choose the right-hand option	n/a
Go cued	chose cued option (e.g. the lit lever)	didn't choose the cued option	n/a
Go uncued	chose the uncued option (e.g. the unlit lever)	didn't choose the uncued option	n/a

Table 2: Exploratory strategy models for rat tasks. Columns give the conditions to define trial t as a success, failure, or null for each exploratory strategy.

	Success	Failure	Null
Win-stay-spatial	Rewarded on trial $t-1$ AND chose the same spatial option (e.g. the left lever) on trial t	Rewarded on trial $t-1$ AND NOT chosen the same spatial option on trial t	Unrewarded on trial $t-1$
Lose-shift-spatial	Unrewarded on trial $t-1$ AND chose a different spatial option (e.g. the left lever) on trial t	Unrewarded on trial $t-1$ AND NOT chosen a different spatial option on trial t	Rewarded on trial $t-1$
Win-stay-cue	Rewarded on trial $t-1$ AND chose the same cued option on trial t (e.g. chose the lit lever on trials $t-1$ and t; or the unlit lever on trials $t-1$ and t)	Rewarded on trial $t-1$ AND NOT chosen the same cued option on trial t	Unrewarded on trial $t-1$
Lose-shift-cue	Unrewarded on trial $t-1$ AND chose a different cued option on trial t from the choice on trial t (e.g. chose the lit lever on trial $t-1$ and the unlit lever on trial t; or the unlit lever on trial $t-1$ and the lit lever on trial t)	Unrewarded on trial $t-1$ AND NOT chosen a different cued option on trial t	Rewarded on trial $t-1$
Alternate	Chose another spatial option compared to the previous trial	Chose the same spatial option as the previous trial	n/a
Sticky	Chose the same spatial option as the previous trial	Chose another spatial option compared to the previous trial	n/a

Table 3: Exploratory strategy models for the non-human primate stimulus-to-action task. Columns give the conditions to define trial t as a success, failure, or null for each exploratory strategy.

	Success	Failure	Null
Repeat-Large-Reward	Obtained the large reward on trial $t-1$ AND chose the same option on trial t	Obtained the large reward on trial $t-1$ AND NOT chosen the same option on trial t	Obtained the small reward on trial $t-1$
Repeat-High-Prob	Chose the higher probability of large reward option on trial $t-1$ AND chose the same option on trial t	Chose the higher probability of large reward option on trial $t-1$ AND NOT chosen the same option on trial t	Chose the lower probability of large reward option on trial $t-1$
Switch-After-Small-Reward	Obtained the small reward on trial $t-1$ AND NOT chosen the same option on trial t	Obtained the small reward on trial $t-1$ AND chosen the same option on trial t	Obtained the large reward on trial $t-1$
Switch-To-High-Prob	Chose the lower probability of large reward option on trial $t-1$ AND NOT chosen the same option on trial t	Chose the lower probability of large reward option on trial $t-1$ AND chose the same option on trial t	Chose the higher probability of large reward option on trial $t-1$.

Interpreting the posterior distributions

The full posterior distribution $P(\text{strategy}_i(t)|\text{choices}(1:t))$ is a Beta distribution with parameters $\alpha_i(t), \beta_i(t)$. In this paper we extract two scalar estimates from that posterior distribution in order to visualise changes in strategies and compare strategies.

First, the best point estimate of probability of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ is its mode, also called its *maximum a posteriori* (MAP) value. We find the MAP value by numerically evaluating the Beta distribution's probability density function, and finding the value of p that corresponds to the peak of the density function.

Second, as a summary of evidence for that MAP probability estimate we compute the precision of the posterior distribution, which is the inverse of the variance of the Beta distribution:

$$\text{Precision} = \left(\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)^{-1}. \tag{2}$$

We then use these scalar estimates as the basis for defining two criterion:

Decision criterion to select the most likely strategy being executed at trial t, choose strategy i with the highest MAP estimate of $P(\text{strategy}_i(t)|\text{choices}(1:t))$ resolving ties by choosing the strategy with the maximum precision.

Learning criterion the first trial at which the MAP probability estimate for the target-rule's strategy remains above chance until the end of the session. If the first trial of a session met this criterion, so all trials were above chance, we chose the trial with the minimum MAP probability estimate, to identify the trial at which the strategy probability increased thereafter.

We apply the decision criterion to the synthetic agent, and suggest it here as a general criterion for situations where the selection of a single dominant strategy per trial is required.

Performance simulations

Synthetic agent on a 2AFC task

To illustrate the Bayesian algorithm, we generated synthetic data that simulated an agent working on a two-alternative forced-choice task, described in Figure 2. The task structure was similar to the tasks described in Figures 1 and 3. The agent could choose between two options (e.g. a right or left lever), and a random cue was alternated between the two options. A reward was delivered only when the agent's choice matched the cue. We generated 500 trials divided into blocks of 100 trials, in which the agent's choice followed a specific strategy ('Go right', 'Alternate','Lose-Shift-Cued', 'Go cued' and 'Lose-Shift-Choice'). We first defined the vector of cue locations as a random binary vector with p=0.5; then we defined the vector of choices. For 'Go right' and 'Alternate' the agent's choice was consistently right or alternating, respectively. For 'Lose-Shift' strategies, we generated a random vector of binary choices; then, for every trial t corresponding to a loss (i.e. not matching the cue), we changed the choice in trial $t+1$ in order to match the cue ('Lose-Shift-Cued') or choice location ('Lose-Shift-Choice'). The 'Go cued' block consisted of choices matching the cue location.

Switching simulations

To explore the performance of the Bayesian algorithm, we ran extensive simulations of the algorithm detecting the switch between two arbitrary strategies. The general model for the switch between the two strategies generated a time-series of the success ($x_i(t) = 1$) or failure ($x_i(t) = 0$) to execute strategy $i \in \{1, 2\}$ on each trial. Strategy 1 was executed on trials 1 to T with probability $p_1 = a$, and strategy 2 was executed on trials 1 to T with probability $p_2 = b$. From trial t the probabilities then reversed linearly:

$$p_1 = \begin{cases} a - m \times j & \text{if } j < (a-b)/m \\ b & \text{otherwise} \end{cases}$$

$$p_2 = \begin{cases} b + m \times j & \text{if } j < (a-b)/m \\ a & \text{otherwise} \end{cases}$$

where j is the number of trials elapsed since T, and m is the rate of change in probability per trial. Figure 2g illustrates the model for the case of $a = 1$ and $b = 0$. Note that the probabilities crossover at $j = (a-b)/2m$; we use this in our definitions of detecting the switch below. From this general model we could explore qualitatively different types of strategy switch: whether the switch was abrupt (m is infinity) or gradual ($m < a - b$).

Given some choice of $\{a,b,T,m\}$, we generated the time-series of x_1 and x_2 for strategies 1 and 2. On every trial t we applied our algorithm to update the two posterior distributions for strategies 1 and 2, and their corresponding MAP estimates $\hat{p}(\text{strategy}_1)$ and $\hat{p}(\text{strategy}_2)$. We defined the detection of the switch as the first trial for which $\hat{p}(\text{strategy}_2) \geq \hat{p}(\text{strategy}_1)$. For abrupt-switching models this was counted from trial T; for gradual-switching models this was counted from the "crossover" trial at which p_2 first exceeded p_1. Stability of detection was defined as the proportion of the next w trials for which this detection criterion remained true; results are presented for $w = 50$.

As shown in Figure 2f, using $\gamma < 1$ causes the detection time to plateau with increasing T, because evidence decay means past history has no effect on detection time. For

further simulations (Figure 2h-i) we thus fixed $T = 500$, to get worst-case performance for detection.

For the detection of a new conditional strategy, we defined a probability per trial $p(\text{condition met}) \in (0, 1]$ that its condition would be met and simulated a model where its probability of execution $p_{\text{new}}(t)$ increased linearly from 0 to 1 at a rate of m per trial. We then created time-series of $x_{\text{new}}(t)$, by first checking if its condition was met, setting $x_{\text{new}}(t) = \text{null}$ if not; if it was met, we then checked if the strategy was executed ($x_{\text{new}}(t) = 1$) or not ($x_{\text{new}}(t) = 0$). We used our algorithm to find the corresponding MAP estimate $\hat{p}(\text{strategy}_{\text{new}})$ from the time-series, and defined detection as the MAP estimate being greater than 0.5. Results are plotted in SI Figure S4.

Effects of missing the true exploratory strategy

A set of tested exploratory strategies may not contain the true exploratory strategy or strategies used by a subject. To examine how the algorithm performs in this case, we determined the proportion of trials on which a given tested strategy would be expected to make the same choice as a true strategy. Formally, this is the conditional probability $p(x_{\text{test}} = 1|\text{true strategy})$ that the tested strategy is a success on any trial given that the true strategy is being used, which we call $p(\text{match})$ in the main text. The algorithm's MAP probability for the tested strategy is an estimate of this conditional probability (SI Figure S5).

To compute $p(x_{\text{test}} = 1|\text{true strategy})$, we decomposed this conditional probability into two main terms, one for trials where the true strategy can be executed, and one for trials where it cannot:

$$p(x_{\text{test}} = 1|\text{true strategy}) = p(x_{\text{test}} = 1|x_{\text{true}} = 1)p(x_{\text{true}} = 1) + p(x_{\text{test}} = 1|x_{\text{true}} = \text{null})p(x_{\text{true}} = \text{null}), \quad (3)$$

with $p(x_{\text{true}} = 1) + p(x_{\text{true}} = \text{null}) = 1$. To understand what happens if the true strategy is missing, we enumerated the possible values of Equation 3 for four cases: both true and tested strategies were deterministic; one was deterministic and the other conditional (giving two cases); and both were conditionally deterministic. We explain the enumerations for these cases below, and plot the results in SI Figure S6.

We first considered the case where the true and tested exploratory strategies were deterministic. In this case $p(x_{\text{true}} = 1) = 1$, and so we needed only to consider the possible values of $p(x_{\text{test}} = 1|x_{\text{true}} = 1)$. For a task with n choices these are $\{0, 1/n, 1 - 1/n, 1\}$: 0, if the tested and true strategy are mutually exclusive (e.g. one repeats the previous choice and one makes a different choice to the previous one); $1/n$, if the true and tested strategy make a single choice; $1 - 1/n$ if the true and tested strategy make anything other than a single choice (e.g. the shift to any other choice than the previous one); 1, if the tested and true strategy are equivalent. The latter implies the true strategy is not missing, so we omitted it; consequently, for the case where both true and test strategies are deterministic, Equation 3 gives just three solutions $p(x_{\text{test}} = 1|\text{true strategy}) = \{0, 1/n, 1-1/n\}$; we plot these three solution as a cumulative distribution in SI Figure S6, panel a. The maximum $p(match)$ is thus $1 - 1/n$ for this case.

For the case where the true strategy is deterministic and the tested strategy is conditional, we first decomposed the probability $p(x_{\text{test}} = 1|x_{\text{true}} = 1)$ into a the probability that the condition for the test strategy is met $p(\text{test condition met})$ and the probability that the

test strategy's choice, if made, matches the true strategy p(test match true—condition met):

$$p(x_{\text{test}} = 1|x_{\text{true}} = 1) = p(\text{test match true — condition met})p(\text{test condition met}). \quad (4)$$

To evaluate Equation 3 using Equation 4, we tested p(test condition met) $\in [0.1, 1)$, and used p(test match true — condition met) $\in \{0, 1/n, 1 - 1/n, 1\}$ as argued for the deterministic case, above. We then evaluated Equation 3 for all combinations of these values in Equation 4, and plot the full set of enumerated probabilities in SI Figure S6, panel b, as a cumulative distribution, omitting the case where the tested and true strategies were entirely equivalent (i.e. $p(x_{\text{test}} = 1|x_{\text{true}} = 1) = 1$). Note that we allowed the possibility of the tested strategy being logically equivalent to the true strategy when the tested strategy's condition was met, by including p(test match true— condition met) $= 1$; if we do not allow this possibility, we get the set of enumerated probabilities plotted in purple in SI Figure S6, panel b.

For the case where the true strategy is conditionally deterministic and the tested strategy is deterministic, we set $p(\text{true} = 1) \in [0.1, 1)$ to model the frequency at which the true strategy's condition was met. We noted that $p(x_{\text{test}} = 1|x_{\text{true}} = 1)$ and $p(x_{\text{test}} = 1|x_{\text{true}} = \text{null})$ were independent and can each take any value in $\{0, 1/n, 1 - 1/n, 1\}$, except both could not be equal to 1 as this would mean the tested and true strategy were logically equivalent. We then evaluated Equation 3 using all combinations of these values for $p(\text{true} = 1)$, $p(x_{\text{test}} = 1|x_{\text{true}} = 1)$, and $p(x_{\text{test}} = 1|x_{\text{true}} = \text{null})$, and plot the enumerated probabilities in SI Figure S6, panel d.

For the case where both the tested and true strategy are conditional, we again used Equation 4 to model the conditional test strategy. We then evaluated Equation 3 using all combinations of values for $p(\text{true} = 1)$ and $p(x_{\text{test}} = 1|x_{\text{true}} = \text{null})$ specified in the preceding paragraph, and for $p(x_{\text{test}} = 1|x_{\text{true}} = 1)$ specified under Equation 4. The resulting enumerated probabilities are plotted in SI Figure S6, panel e.

In reality we only get to specify the type of test strategy and do not know the type (conditional or deterministic) of true strategy adopted by the subject. Thus, to put bounds on what we can infer from the observed MAP probabilities for a given type of test strategy, we combined the enumerated probabilities for the deterministic and conditional true strategies. The resulting distributions are plotted in the last column of SI Figure S6.

Derivation of the iterative update and evidence decay

We first rehearse the standard Bayesian approach to estimating a binomial probability p as a random variable when that probability is stationary. We then introduce the decay of evidence to track non-stationary probabilities, and obtain the iterative update rules given in the main text.

Our data are a sequence of Bernoulli trials $\tau = 1, 2, \ldots t$, each of which is either a success $x(\tau) = 1$ or a failure $x(\tau) = 0$ of an event. A running total of the successful events is $s(t) = \sum_{\tau=1}^{t} x(\tau)$. The posterior distribution $\pi(p|s)$ for estimating the probability of that number of successful events is then given by Bayes theorem

$$\pi(p|s) = \frac{P(s|p)\pi(p)}{\int_0^1 P(s|p)\pi(p)dp}, \quad (5)$$

where the denominator is the normalisation constant to obtain a probability distribution function. (Note we write $\pi(p|s)$ as $P(\text{strategy}_i|\text{choices}(1 : t))$ in the main text when applied to our specific problem of estimating the probability of strategy i).

As noted in the main text, the likelihood $P(s|p)$ is the binomial distribution and we chose the prior $\pi(p)$ to be the Beta distribution with parameters α, β. Using those in Equation (5) gives

$$\pi(p|s) = \frac{1}{Z} \underbrace{\left[\binom{t}{s} p^s (1-p)^{t-s}\right]}_{\text{likelihood}} \underbrace{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]}_{\text{prior}}, \tag{6}$$

where Γ is the gamma function and Z is the denominator in Equation 5.

The solution of Equation (6) is another Beta distribution:

$$\pi(p|s) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha+s-1}(1-p)^{\beta+t-s-1} \tag{7}$$

which is equivalent to simply updating the Beta distribution parameters as $(\alpha(0)+s, \beta(0)+t-s)$ given some starting values of $\alpha(0), \beta(0)$, which define the prior distribution.

Rather than updating after t trials have elapsed, we can instead update every trial by using the posterior of the previous trial $\pi(p|s(t-1))$ as the prior for the update on the current trial t. To see this, we can rewrite Equation (5) to update on every trial t by substituting the Bernoulli distribution $p^x(1-p)^{(1-x)}$ for the binomial distribution, and obtain the iterative equivalent of Equation (7):

$$\pi(p|s(t)) = Bp^{\alpha(t-1)+x-1}(1-p)^{\beta(t-1)-x}, \tag{8}$$

where B is the constant $\frac{\Gamma(\alpha(t-1)+\beta(t-1))}{\Gamma(\alpha(t-1))\Gamma(\beta(t-1))}$.

Equation (8) is equivalent to updating the parameters of the Beta distribution on every trial t as $(\alpha(t-1)+x(t), \beta(t-1)+(1-x(t)))$, given some prior values for $\alpha(0), \beta(0)$:

$$\begin{array}{lll} \text{Trial 1} & \alpha(1) = \alpha(0) + x(1), & \beta(1) = \beta(0) + (1 - x(1)) \\ \text{Trial 2} & \alpha(2) = \alpha(1) + x(2), & \beta(2) = \beta(1) + (1 - x(2)) \\ & \vdots & \vdots \\ \text{Trial } t & \alpha(t) = \alpha(t-1) + x(t), & \beta(t) = \beta(t-1) + (1 - x(t)). \end{array} \tag{9}$$

It is this iterative form that allows computationally efficient real-time updates of the probability of each strategy.

We now turn to our solution to the problem that p is not stationary when estimating the use of strategies. Our idea is to decay the evidence entering the posterior update in Equation (5). Consider the following models for the exponential decay of the Bernoulli trial outcomes up to trial t, for the successes

$$s^*(t) = \sum_{\tau=1}^{t} \gamma^{t-\tau} x(\tau), \tag{10}$$

and the failures

$$f^*(t) = \sum_{\tau=1}^{t} \gamma^{t-\tau} (1 - x(\tau)), \tag{11}$$

where $\gamma \in (0, 1]$ is the rate of evidence decay.

We can substitute s^* for s and f^* for $t-s$ in Equation (6) and obtain the posterior distribution

$$\pi(p|s^*(t), f^*(t)) = Bp^{\alpha+s^*(t)-1}(1-p)^{\beta+f^*(t)-1}, \tag{12}$$

for the estimate of p on trial t given the decayed evidence up to that trial. Again, this is equivalent to updating the Beta distribution parameters as $(\alpha(0)+s^*(t), \beta(0)+f^*(t))$.

To obtain the trial-by-trial update equivalent to Equation (12), we note that the iterative forms of Equations (10) and (11) are

$$\begin{aligned} s^*(t) &= \gamma s^*(t-1) + x(t), \\ f^*(t) &= \gamma f^*(t-1) + (1 - x(t)), \end{aligned} \tag{13}$$

starting from $s^*(0) = f^*(0) = 0$.

With these the update of the Beta distribution parameters on trial t becomes

$$\begin{aligned} \alpha(t) &= \alpha(0) + \gamma s^*(t-1) + x(t) \\ \beta(t) &= \beta(0) + \gamma f^*(t-1) + (1 - x(t)). \end{aligned} \tag{14}$$

Together, Equations (13) and Equations (14) form the algorithm we use here.

One consideration when setting the decay parameter γ is that it creates asymptotic limits for the values of α and β. Equations (10) and (11) are geometric series, so the asymptotic values of α and β have exact solutions. For a long run of successes where $x(t) = 1$ as $t \to \infty$ these are

$$\alpha \to \alpha(0) + \frac{1}{1-\gamma}, \quad \beta \to \beta(0). \tag{15}$$

Similarly, for a long run of failures where $x(t) = 0$ as $t \to \infty$ these are

$$\alpha \to \alpha(0), \quad \beta \to \beta(0) + \frac{1}{1-\gamma}. \tag{16}$$

These asymptotic limits have two consequences. The first is that choosing $\gamma \ll 1$ will lead to large changes in $P(\text{strategy}_i(t)|\text{choices}(1:t))$ on every update of α or β, and so also of the MAP probability estimate (as we demonstrate in SI Figure S3 for the synthetic agent). The second is that the MAP probability estimate may only approach 1 slowly, so we do not use this as a criterion.

If we use $\gamma < 1$ then the precision has asymptotic limits found by substituting Equations (15) or (16) into Equation (2), as appropriate.

References

Akrami, A., Kopec, C. D., Diamond, M. E., and Brody, C. D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature*, 554:368–372.

Amarante, L. M., Caetano, M. S., and Laubach, M. (2017). Medial frontal theta is entrained to rewarded actions. *J Neurosci*, 37:10757–10769.

Ashwood, Z. C., Roy, N. A., Stone, I. R., Laboratory, I. B., Urai, A. E., Churchland, A. K., Pouget, A., and Pillow, J. W. (2022). Mice alternate between discrete strategies during perceptual decision-making. *Nat Neurosci*, 25:201–212.

Banerjee, A., Parente, G., Teutsch, J., Lewis, C., Voigt, F. F., and Helmchen, F. (2020). Value-guided remapping of sensory cortex by lateral orbitofrontal cortex. *Nature*, 585:245–250.

Bartolo, R. and Averbeck, B. B. (2020). Prefrontal cortex predicts state switches during reversal learning. *Neuron*, 106:1044–1054.e4.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nat Neurosci*, 10:1214–1221.

Birrell, J. M. and Brown, V. J. (2000). Medial frontal cortex mediates perceptual attentional set shifting in the rat. *J Neurosci*, 20:4320–4324.

Boulougouris, V., Dalley, J. W., and Robbins, T. W. (2007). Effects of orbitofrontal, infralimbic and prelimbic cortical lesions on serial spatial reversal learning in the rat. *Behavioural brain research*, 179:219–228.

Brady, A. M. and Floresco, S. B. (2015). Operant procedures for assessing behavioral flexibility in rats. *Journal of visualized experiments : JoVE*, page e52387.

Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J Neurosci*, 12(12):4745–4765.

Buckley, M. J., Mansouri, F. A., Hoda, H., Mahboubi, M., Browning, P. G. F., Kwok, S. C., Phillips, A., and Tanaka, K. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science*, 325:52–58.

Campagner, D., Evans, M. H., Chlebikova, K., Colins-Rodriguez, A., Loft, M. S. E., Fox, S., Pettifer, D., Humphries, M. D., Svoboda, K., and Petersen, R. S. (2019). Prediction of choice from competing mechanosensory and choice-memory cues during active tactile decision making. *J Neurosci*, 39:3921–3933.

Carandini, M. and Churchland, A. K. (2013). Probing perceptual decisions in rodents. *Nature neuroscience*, 16:824–831.

Churchland, A. K., Kiani, R., and Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature neuroscience*, 11:693–702.

Constantinople, C. M., Piet, A. T., Bibawi, P., Akrami, A., Kopec, C., and Brody, C. D. (2019). Lateral orbitofrontal cortex promotes trial-by-trial learning of risky, but not spatial, biases. *eLife*, 8.

Donoso, M., Collins, A. G. E., and Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, 344:1481–1486.

Durstewitz, D., Vittoz, N. M., Floresco, S. B., and Seamans, J. K. (2010). Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, 66:438–448.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn Sci*, 14(3):119–130.

Floresco, S. B., Block, A. E., and Tse, M. T. L. (2008). Inactivation of the medial prefrontal cortex of the rat impairs strategy set-shifting, but not reversal learning, using a novel, automated procedure. *Behav Brain Res*, 190:85–96.

Gallistel, C. R., Fairhurst, S., and Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proc Natl Acad Sci U S A*, 101(36):13124–13131.

Genovesio, A., Brasted, P. J., Mitz, A. R., and Wise, S. P. (2005). Prefrontal cortex activity related to abstract response strategies. *Neuron*, 47:307–320.

Giurfa, M. and Sandoz, J.-C. (2012). Invertebrate learning and memory: Fifty years of olfactory conditioning of the proboscis extension response in honeybees. *Learning & Memory*, 19:54–66.

Guise, K. G. and Shapiro, M. L. (2017). Medial prefrontal cortex reduces memory interference by modifying hippocampal encoding. *Neuron*, 94:183–192.e8.

Hanks, T. D. and Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. *Neuron*, 93:15–31.

Harris, C., Aguirre, C., Kolli, S., Das, K., Izquierdo, A., and Soltani, A. (2021). Unique features of stimulus-based probabilistic reversal learning. *Behavioral neuroscience*, 135:550–570.

Honegger, K. and de Bivort, B. (2018). Stochasticity, individuality and behavior. *Curr Biol*, 28:R8–R12.

International Brain Laboratory, Aguillon-Rodriguez, V., Angelaki, D., Bayer, H., Bonacchi, N., Carandini, M., Cazettes, F., Chapuis, G., Churchland, A. K., Dan, Y., Dewitt, E., Faulkner, M., Forrest, H., Haetzel, L., Häusser, M., Hofer, S. B., Hu, F., Khanal, A., Krasniak, C., Laranjeira, I., Mainen, Z. F., Meijer, G., Miska, N. J., Mrsic-Flogel, T. D., Murakami, M., Noel, J.-P., Pan-Vazquez, A., Rossant, C., Sanders, J., Socha, K., Terry, R., Urai, A. E., Vergara, H., Wells, M., Wilson, C. J., Witten, I. B., Wool, L. E., and Zador, A. M. (2021). Standardized and reproducible measurement of decision-making in mice. *eLife*, 10.

Ito, H. T., Zhang, S.-J., Witter, M. P., Moser, E. I., and Moser, M.-B. (2015). A prefrontal-thalamo-hippocampal circuit for goal-directed spatial navigation. *Nature*, 522:50–55.

Izquierdo, A., Brigman, J. L., Radke, A. K., Rudebeck, P. H., and Holmes, A. (2017). The neural basis of reversal learning: An updated perspective. *Neuroscience*, 345:12–26.

Jang, A. I., Costa, V. D., Rudebeck, P. H., Chudasama, Y., Murray, E. A., and Averbeck, B. B. (2015). The role of frontal cortical and medial-temporal lobe brain areas in learning a bayesian prior belief on reversals. *J Neurosci*, 35:11751–11760.

Juavinett, A. L., Erlich, J. C., and Churchland, A. K. (2017). Decision-making behaviors: weighing ethology, complexity, and sensorimotor compatibility. *Curr Opin Neurobiol*, 49:42–50.

Karlsson, M. P., Tervo, D. G. R., and Karpova, A. Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, 338:135–139.

Kim, J.-N. and Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, 2(2):176–185.

Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*, 27(12):712–719.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93:480–490.

Lau, B. and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, 84:555–579.

Leeson, V. C., Robbins, T. W., Matheson, E., Hutton, S. B., Ron, M. A., Barnes, T. R. E., and Joyce, E. M. (2009). Discrimination learning, reversal, and set-shifting in first-episode schizophrenia: stability over six years and specific associations with medication type and disorganization syndrome. *Biol Psychiatry*, 66:586–593.

Mansouri, F. A., Freedman, D. J., and Buckley, M. J. (2020). Emergence of abstract rules in the primate brain. *Nat Rev Neurosci*, 21:595–610.

Miller, K. J., Botvinick, M. M., and Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature neuroscience*, 20:1269–1276.

Packard, M. G. and McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of learning and memory*, 65:65–72.

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J Vis*, 5(5):376–404.

Pereira, T. D., Shaevitz, J. W., and Murthy, M. (2020). Quantifying behavior to understand the brain. *Nature Neurosci*, 23:1537–1549.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442:1042–1045.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12:919–926.

Pezze, M., McGarrity, S., Mason, R., Fone, K. C., and Bast, T. (2014). Too little and too much: hypoactivation and disinhibition of medial prefrontal cortex cause attentional deficits. *J Neurosci*, 34:7931–7946.

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat Neurosci*, 16(9):1170–1178.

Powell, N. J. and Redish, A. D. (2016). Representational changes of latent strategies in rat medial prefrontal cortex precede changes in behaviour. *Nature communications*, 7:12830.

Raposo, D., Sheppard, J. P., Schrater, P. R., and Churchland, A. K. (2012). Multisensory decision-making in rats and humans. *J Neurosci*, 32:3726–3735.

Rich, E. L. and Shapiro, M. (2009). Rat prefrontal cortical neurons selectively code strategy switches. *J Neurosci*, 29(22):7208–7219.

Rich, E. L. and Shapiro, M. L. (2007). Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks. *J Neurosci*, 27(17):4747–4755.

Roy, N. A., Bak, J. H., Laboratory, I. B., Akrami, A., Brody, C. D., and Pillow, J. W. (2021). Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron*, 109:597–610.e6.

Rudebeck, P. H., Behrens, T. E., Kennerley, S. W., Baxter, M. G., Buckley, M. J., Walton, M. E., and Rushworth, M. F. S. (2008). Frontal cortex subregions play distinct roles in choices between actions and stimuli. *J Neurosci*, 28:13775–13785.

Russo, E., Ma, T., Spanagel, R., Durstewitz, D., Toutounji, H., and Köhr, G. (2021). Coordinated prefrontal state transition leads extinction of reward-seeking behaviors. *J Neurosci*, 41:2406–2419.

Shiner, T., Symmonds, M., Guitart-Masip, M., Fleming, S. M., Friston, K. J., and Dolan, R. J. (2015). Dopamine, salience, and response set shifting in prefrontal cortex. *Cerebral Cortex*, 25:3629–3639.

Singh, A., Peyrache, A., and Humphries, M. D. (2019). Medial prefrontal cortex population activity is plastic irrespective of learning. *J Neurosci*, 39:3470–3483.

Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A. M., Suzuki, W. A., and Brown, E. N. (2004). Dynamic analysis of learning in behavioral experiments. *J Neurosci*, 24(2):447–461.

Sohn, H., Narain, D., Meirhaeghe, N., and Jazayeri, M. (2019). Bayesian computation through cortical latent dynamics. *Neuron*, 103:934–947.e5.

Suzuki, W. A. and Brown, E. N. (2005). Behavioral and neurophysiological analyses of dynamic learning processes. *Behavioral and cognitive neuroscience reviews*, 4:67–95.

Tait, D. S., Bowman, E. M., Neuwirth, L. S., and Brown, V. J. (2018). Assessment of intradimensional/extradimensional attentional set-shifting in rats. *Neuroscience and biobehavioral reviews*, 89:72–84.

Trepka, E., Spitmaan, M., Bari, B. A., Costa, V. D., Cohen, J. Y., and Soltani, A. (2021). Entropy-based metrics for predicting choice behavior based on local response to reward. *Nature communications*, 12:6567.

van Wingerden, M., Vinck, M., Lankelma, J., and Pennartz, C. M. A. (2010). Theta-band phase locking of orbitofrontal neurons during reward expectancy. *J Neurosci*, 30:7078–7087.

Wang, J., Tait, D. S., Brown, V. J., and Bowman, E. M. (2019). Exacerbation of the credit assignment problem in rats with lesions of the medial prefrontal cortex is revealed by bayesian analysis of behavior in the pre-solution period of learning. *Behavioural brain research*, 372:112037.

Wu, Z., Kavanova, M., Hickman, L., Boschin, E. A., Galeazzi, J. M., Verhagen, L., Ainsworth, M., Pedreira, C., and Buckley, M. J. (2021). Low-beta repetitive transcranial magnetic stimulation to human dorsolateral prefrontal cortex during object recognition memory sample presentation, at a task-related frequency observed in local field potentials in homologous macaque cortex, impairs subsequent recollection but not familiarity. *Eur J Neurosci*, 54:7918–7945.

Yin, H. H. and Knowlton, B. J. (2004). Contributions of striatal subregions to place and response learning. *Learn Mem*, 11(4):459–463.

Tracking subjects' strategies in behavioural choice experiments at trial resolution

Silvia Maggi, Rebecca M. Hock, Martin O'Neill, Mark J. Buckley,
Paula Moran, Tobias Bast, Musa Sami, and Mark D. Humphries

Supplementary Information

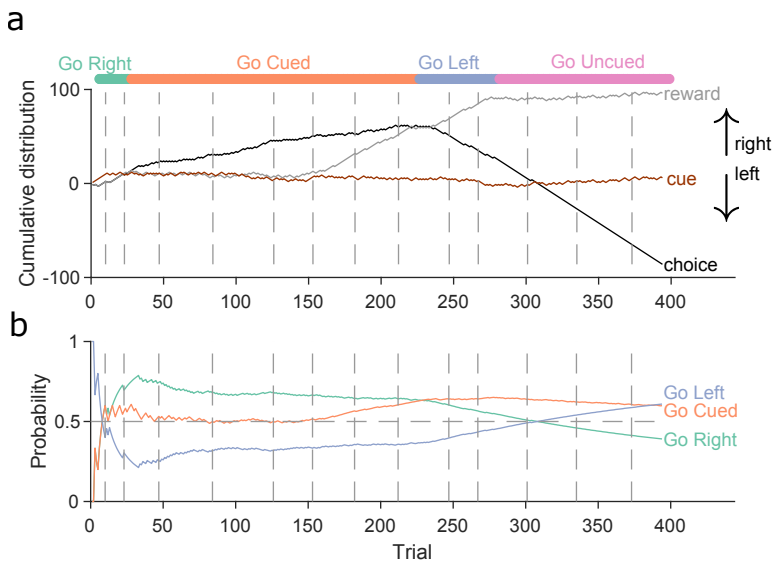

Figure S1: A stationary Bayesian approach fails to track behavioural changes.
(**a**) Same as Figure 1b. Example rat performance on the Y-maze task. Each curve show the cumulative distribution of reward (grey), choice (black) and cue location (brown). Vertical grey dashed lines indicate sessions. Colours at the top indicate the target rule.
(**b**) MAP probabilities for three strategies across the 14 sessions for the example animal in panel a, when using Bayesian estimation without evidence decay. Note how the estimated probability of 'go right' remains the highest through both the switches to the 'go cued' and 'go left' rules, despite the animal clearly shifting its behaviour, as it required ten consecutive trials (or 11 out of 12) of the correct choice to switch the rule. Horizontal grey dashed line indicates chance level.

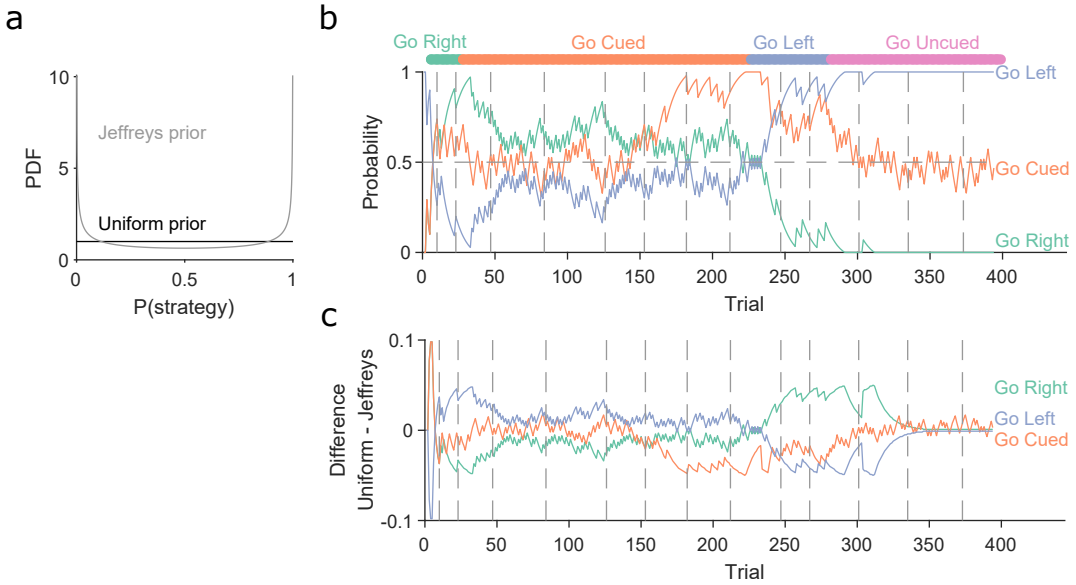

Figure S2: The Bayesian approach is robust to changes in the initial prior.
(**a**) Example of uniform and Jeffrey's distributions used as the prior probability for the Bayesian algorithm.
(**b**) Similar to Figure 1d, MAP probabilities over trials for the example rat in Figure 1b. The model is initialized with the Jeffrey's prior.
(**c**) Difference between the MAP probabilities for each strategy when initialized with either the uniform or the Jeffrey's prior.

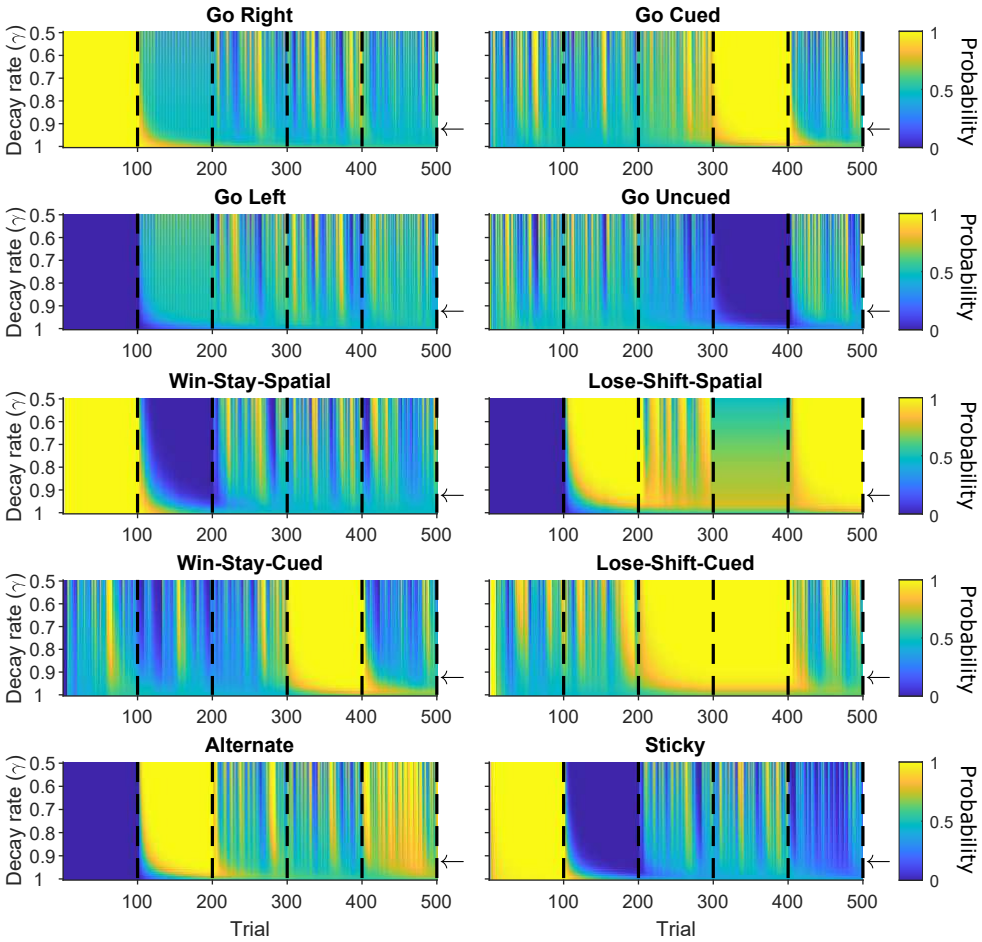

Figure S3: Robust MAP estimation for changes in decay rate.
For the synthetic agent performing the five blocks of strategies, each panel plots the MAP values for one strategy, across each trial and each value of the decay rate between 0.5 and 1, in steps of 0.01. Vertical black dashed lines indicate block switches. Horizontal arrows are $\gamma = 0.9$.

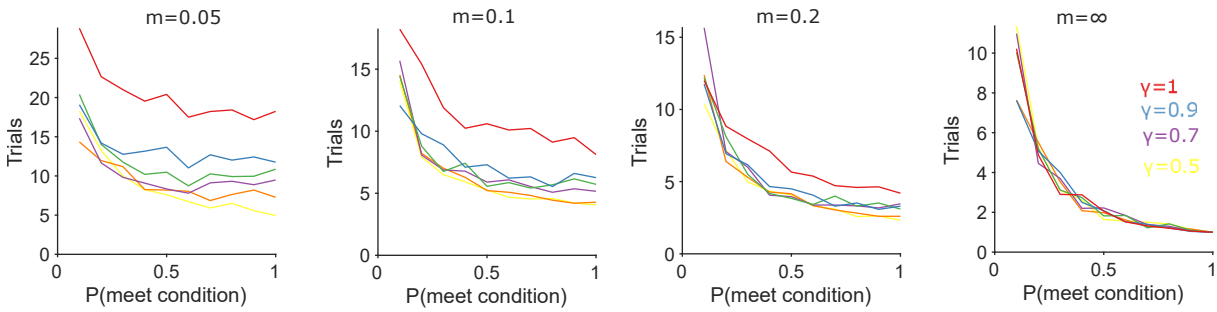

Figure S4: Algorithm performance for detecting a new conditional strategy.
Each panel plots the number of trials needed to detect a new conditional strategy whose condition occurs with probability p(meet condition). The probability of executing the new conditional strategy linearly increases from 0 to 1 with rate m. The right-most panel is the deterministic case where the strategy's probability is immediately 1.

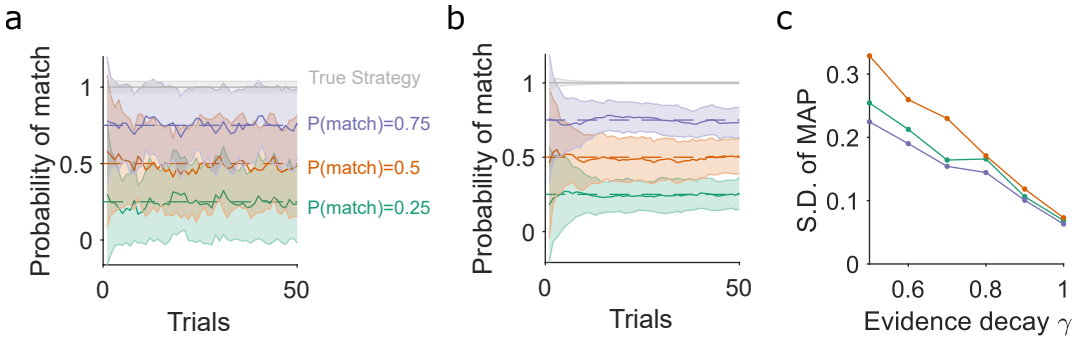

Figure S5: MAP probability converges on the match between a true and tested strategy.
(**a**) Estimating the probability that a tested strategy matches an agent's true strategy. For three different probabilities of matching (p(match)), we plot the mean and standard deviation (S.D.) of the algorithm's MAP estimate of that test strategy, over 50 simulations of 50 trials. Results shown for $\gamma = 0.5$.
(**b**) As for (a), for $\gamma = 0.9$.
(**c**) Variation in the MAP estimates of p(match) decreases with lower rates of evidence decay (higher γ). S.D. taken from trial 50 of 50 simulations similar to those in panels a and b.

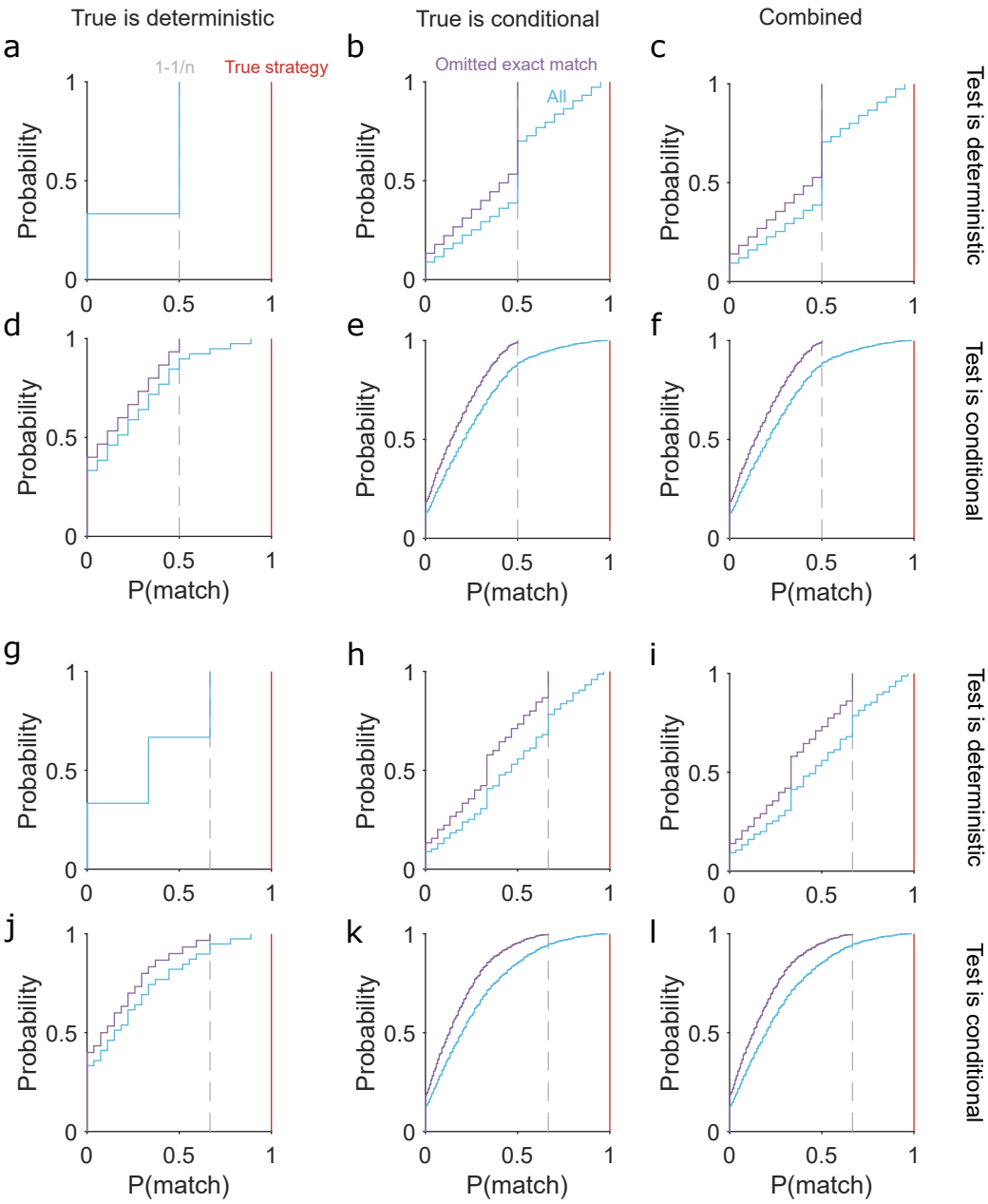

Figure S6: Limits on the probability of a tested strategy matching the true strategy used by an agent during exploration.
(**a**) The cumulative distribution of p(match) between the tested and true strategy when both are deterministic. For a task with n options p(match) can take one of three values: 0 (tested and true are mutually exclusive); $1/n$ (tested strategy chooses one option); $1 - 1/n$ (the tested strategy chooses any but one option).
(**b**) The distribution of p(match) when the tested strategy is deterministic, but the true strategy is conditionally deterministic (blue: "All"). We evaluated p(match) using probabilities between 0.1 and 1 that the true strategy's condition is met; and we allowed that the tested strategy could exactly match the true strategy either when the condition is met or when it is not. Not allowing this exact matching gives the distribution in purple, with a maximum at $1 - 1/n$.
(**c**) The full distribution of p(match) for when the tested strategy is deterministic, given by the combination of the relevant distributions in panels (a) and (b). Because we know the type of tested strategy, but not the type of true strategy used by a subject, these distributions place bounds on what we can infer from the MAP probability of a tested deterministic strategy.
(**d**)–(**f**) As for (a)-(c), for when the tested strategy is conditionally deterministic.
(**g**)–(**l**) As for (a)-(f), for a task with n=3 options. Note the maximum p(match) at $1 - 1/n$ (vertical dashed grey line) when the logical equivalence between tested and true strategies is omitted.

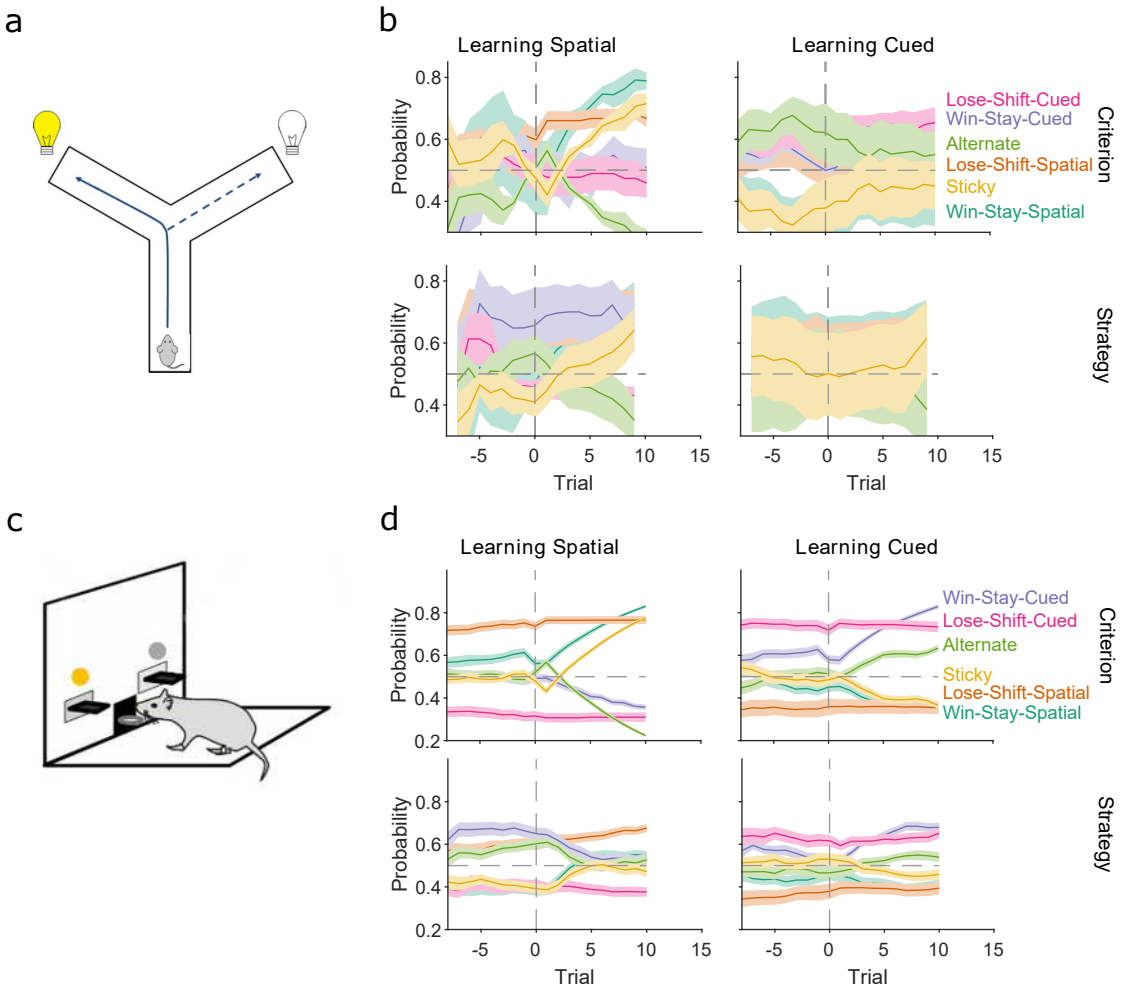

Figure S7: Exploratory strategies around learning
(**a**) Graphical representation of the Y-maze task.
(**b**) MAP probabilities for all tested exploratory strategies aligned to the learning of spatial (left panel) and cued (right panel) rules (trial 0, vertical grey dashed line). Horizontal grey dashed lines are the chance level. Top: aligned to the learning trial found by the study's original learning criterion. Bottom: aligned to the learning trial found by our strategy criterion. Average (bold line) ± SEM (shaded area). N as per Figure 5a, main text.
(**c**) Graphical representation of the lever-press task.
(**d**) Same as panel b for the lever-press task. N as per Figure 5d, main text.

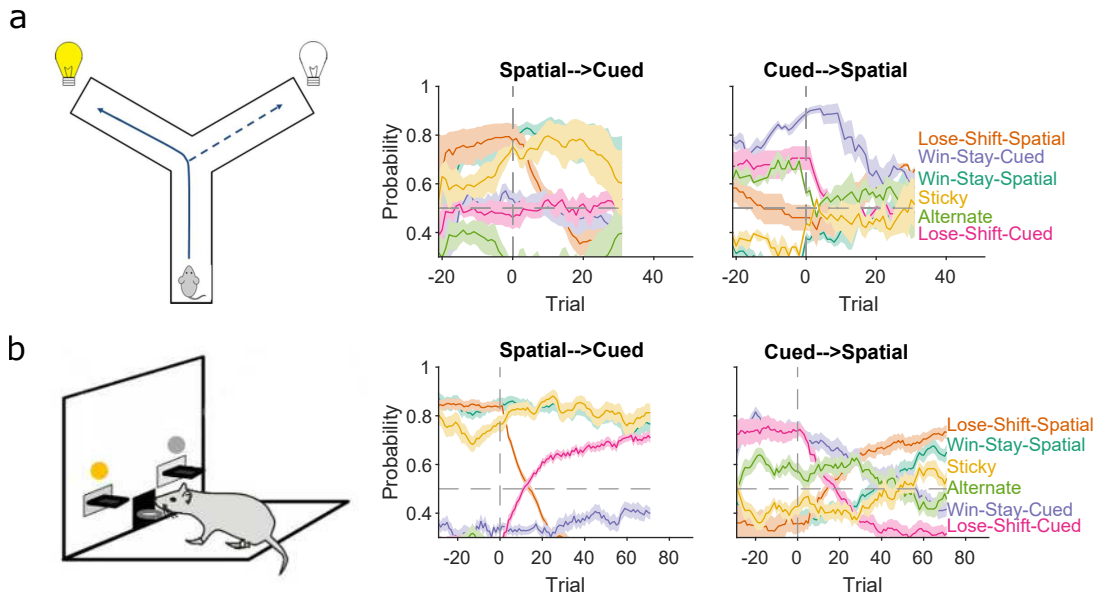

Figure S8: Exploratory strategies around rule switches
(**a**) Responses to rule switches in Y-maze task. MAP probabilities for all tested exploratory strategies aligned around the rule change trial (trial 0) for switches from spatial to cued rules (N=5 switches, left panel) and from cued to spatial rules (N=3 switches, right panel). Curves plot mean ± SEM (shaded area).
(**b**) Same as panel a, for the lever-press task (N=16 animals for spatial→cued; N=12 cued→spatial).

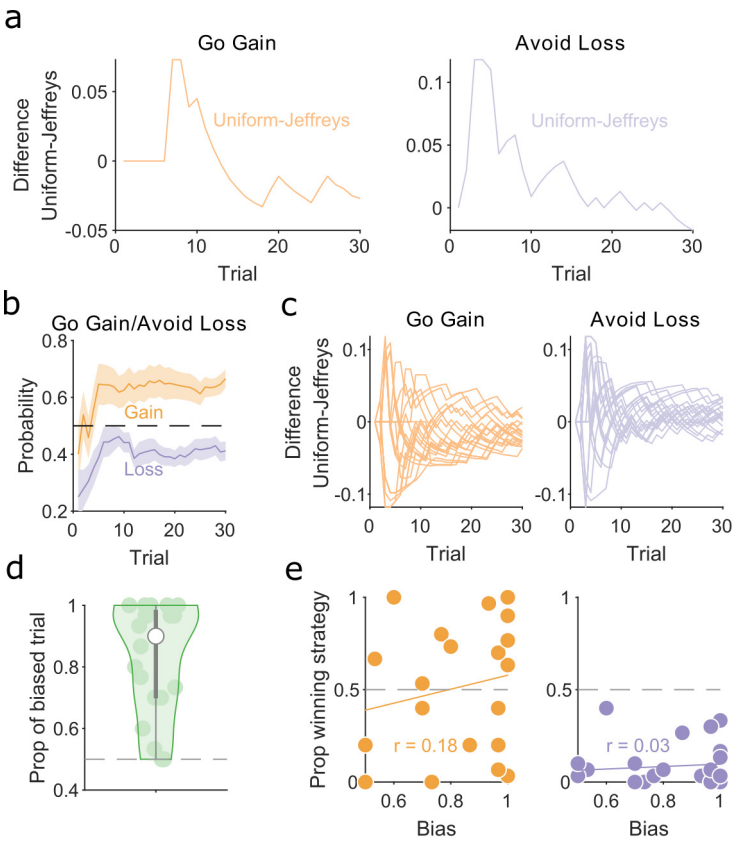

Figure S9: Human strategies are robust to priors and do not correlate with bias in preferred target locations.
(**a**) Difference between MAP probability of 'go gain' and 'avoid loss' strategies initialized with uniform prior and Jeffrey's prior for the example participant in Figure 6b.
(**b**) MAP probability profiles for 'go gain' and 'avoid loss' strategies along the session during gain and loss trials, respectively. Mean (solid curve) ± SEM (shaded area) across participants (N=20). Horizontal dashed line indicates chance level. These strategies were initialized with Jeffrey's prior.
(**c**) Similar to panel a, the difference between the MAP probability of 'go gain' and 'avoid loss' strategies initialized with uniform and Jeffrey's prior for every participant (one line per participant).
(**d**) The distribution of biased trials during Look trials, defined as the proportion of trials in which the MAP probability for using a direction strategy towards one of the two stimulus locations was consistently above chance. No bias corresponds to 0.5, whereas 1 is when the participant chose the same location on every Look trial. Each dot is a participant. White dot is median and grey bars show the 25th and 75th percentiles.
(**e**) The proportion of trials in which 'go gain' (left panel) and 'avoid loss' (right panel) were the winning strategy for each subject as function of their look bias. Each dot is a participant. r is the Pearson correlation coefficient. Solid line plots the linear fit to the data.