Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task

Lilla Horvath[a], Stanley Colcombe[b], Michael Milham[b], Shruti Ray[c], Philipp Schwartenbeck[d], Dirk Ostwald[a]

[a] *Computational Cognitive Neuroscience, Freie Universität Berlin*
[b] *Center for Biomedical Imaging and Neuromodulation, Nathan Kline Institute, Orangeburg*
[c] *Department of Biomedical Engineering, New Jersey Institute of Technology*
[d] *Wellcome Centre for Integrative Neuroimaging, University of Oxford, & Wellcome Trust Centre for Human Neuroimaging, University College London*

Abstract

Humans often face sequential decision-making problems, in which information about the environmental reward structure is detached from rewards for a subset of actions. In the current exploratory study, we introduce an information-selective symmetric reversal bandit task to model such situations and obtained choice data on this task from 24 participants. To arbitrate between different decision-making strategies that participants may use on this task, we developed a set of probabilistic agent-based behavioral models, including exploitative and explorative Bayesian agents, as well as heuristic control agents. Upon validating the model and parameter recovery properties of our model set and summarizing the participants' choice data in a descriptive way, we used a maximum likelihood approach to evaluate the participants' choice data from the perspective of our model set. In brief, we provide quantitative evidence that participants employ a belief state-based hybrid explorative-exploitative strategy on the information-selective symmetric reversal bandit task, lending further support to the finding that humans are guided by their subjective uncertainty when solving exploration-exploitation dilemmas.

Keywords: bandit problem, agent-based behavioural modelling, exploration, exploitation

Introduction

Uncertainty is an inherent part of real-life sequential decision making. Humans often face new and changing situations without being able to directly observe the statistical regularities of the environmental reward structure. Consequently, in their quest to maximize their cumulative rewards, humans have to alternate between exploration and exploitation. Exploration refers to decisions that maximize information gain and thus reduce the uncertainty about the statistical regularities of the environment. Exploitation refers to decisions that maximize reward gain by harnessing the accumulated knowledge.

A standard behavioural test bed to study sequential decision making under uncertainty is the bandit paradigm (e.g. Robbins, 1952; Brand et al., 1956; Brand & Woods, 1957; Berry & Fristedt, 1985; Cohen et al., 2007; Even-Dar et al., 2006; Dayan & Daw, 2008; Bubeck et al., 2009; Gabillon et al., 2012). Two variants of the bandit paradigm have been widely adopted to model real-life sequential decision making under uncertainty. We here refer to these variants as the *classical bandit paradigm* (by some also referred to as partial-feedback paradigm (c.f. Hertwig, 2012; Wulff et al., 2018)) and the *pure exploration paradigm* (by some also referred to as sampling paradigm, ibid.). In both variants, on each trial the deciding agent has to

choose among a finite set of actions with different expected reward values and subsequently observes a reward with probability specific to the chosen action. While the actions' expected rewards are not directly observable, the agent can estimate them by integrating information over multiple reward observations. The difference between the classical bandit paradigm and the pure exploration paradigm stems from the respective challenges they pose for the deciding agent. In the classical bandit paradigm, the agent's task is to maximize the cumulative reward across all trials, while the reward observation confers both information and reward on each trial. The classical bandit paradigm thus raises the problem of how to strike a balance between exploration and exploitation on each trial. In contrast, in the pure exploration paradigm, the aim is to maximize the reward obtained on a single final trial. Here, reward observations on preceding trials only confer information, but not reward. The number of trials preceding this final trial is self-determined by the agent. The pure exploration paradigm thus raises the problem of how to strike a balance between the exploration costs preceding the final trial and the potential final trial reward (Ostwald et al., 2015).

While a large variety of real-life decision-making problems can be modelled with the classical bandit and the pure exploration paradigms, neither variant is suited to model a class of decision-making problems, in which each available action yields positive or negative reward, while only some actions also yield information about the problem's reward structure. Consider for example the situation of a patient who exhibits Corona symptoms during the global pandemic. Assuming that reliable and medically administered Corona tests are available, while self-administered Corona tests are not (as has been the case in many countries in the first year of the pandemic), the patient faces the following decision dilemma: on the one hand, the patient may home-quarantine, thereby minimizing the risk for virus transmission, but at the price of not obtaining information about their own state of infection. Alternatively, the patient may use public transport to undergo a medically administered Corona test, thereby incurring the risk of further transmitting the virus, but obtaining reliable information about their personal infectious state. Situations of this type are similar to the ones modelled with the classical bandit paradigm as each action has a positively or negatively rewarded consequence (which in the example is of societal nature). Importantly, however, in this situation reward-relevant information is detached from reward for one of the actions (home-quarantining), akin to the pure exploration paradigm. Consequently, such a decision situation poses a more pronounced exploration-exploitation dilemma then both classical bandit paradigms and the pure exploration paradigm, because the decision maker is forced to explicitly evaluate the benefit of information gain against the benefit of reward gain. The aim of the current study is to computationally characterize human sequential decision making in such problems.

To this end, we introduce an information-selective symmetric reversal bandit task, which shares key characteristics with the classical symmetric two-armed reversal bandit task (e.g. Bartolo & Averbeck, 2020; Costa et al., 2016; Gläscher et al., 2009; Hauser et al., 2014), but in which information is randomly withheld for either the action with the high or the low expected reward value. To arbitrate between different sequential decision-making strategies that humans may employ on this task, we formulate a set of agent-based behavioural models. We here follow up on recent results showing that one way humans balance between exploration and exploitation is to add an 'information bonus' to the value estimate of an action, which reflects the associated uncertainty (e.g. Gershman, 2018, 2019; Lee et al., 2011; Wilson et al., 2014; Wu et al., 2018). More specifically, we formulate Bayesian agents that represent subjective uncertainty about the structure of the environment in the form of a belief state. The Bayesian agents use the belief state to make either exploitative (i.e., value estimate maximizing actions), explorative (i.e., information bonus maximizing actions), or hybrid explorative-exploitative (i.e., combined value

estimate and information bonus maximizing) actions. Notably, we adopt a Bayesian treatment of exploration and quantify the information bonus as the expected Bayesian surprise (Itti & Baldi, 2009; Sun et al., 2011; Ostwald et al., 2012). In addition to the Bayesian agents, we also formulate belief state-free agents that implement simple strategies, such as a cognitive null model and the win-stay-lose-switch heuristic (Robbins, 1952). Upon validating our modelling initiative, we provide evidence for a belief state-based hybrid explorative-exploitative strategy based on choice data from 24 participants. In summary, we show that in a scenario where every decision has an economic consequence, but only some decisions are informative about the statistical reward structure of the environment, humans are guided by their subjective uncertainty when resolving the exploration-exploitation dilemma.

Methods

Experimental methods

Participants Young adults were recruited from the Nathan Kline Institute Rockland Sample (NKI-RS), a community-ascertained and comprehensively characterized participant sample of more than 1000 individuals between 6 and 85 years of age (Nooner et al., 2012). We initially intended to enroll individuals from the lower and upper ends of the attention deficit hyperactivity disorder (ADHD) spectrum because we were interested in the relationship between ADHD symptoms and behavioural strategies in our task. Yet, the final sample of 24 individuals (12 female, 23 right-handed, age range: 18-35 years, mean age: 24.5 years, standard deviation age: 5.5 years) represented the mid-range of the ADHD spectrum. Moreover, individuals were only invited if they had no lifetime history of severe neurological or psychiatric disorder. We therefore treated the group of participants as a healthy sample and did not conduct analyses to relate ADHD symptoms to task behavior. For additional details about the recruitment and sample characteristics, please refer to Section S.1: Sample characteristics.

Procedure The study consisted of a one-time visit of 3.5 hours to the Nathan Kline Institute for Psychiatric Research (Orangeburg, NY, US). After providing written informed consent, participants were first requested to fill out a series of questionnaires measuring symptoms of ADHD and other mental disorders. Next, participants received detailed written instructions about the information-selective symmetric reversal bandit task and were encouraged to ask any clarification questions. For the detailed instructions provided to the participants, please refer to Section S.2: Participant instructions. To familiarize participants with the task, they next completed a test run of the task on a desktop computer. Finally, participants completed two experimental task runs in a Magnetic Resonance Imaging (MRI) scanner, while behavioural, eye tracking, and functional MRI data were acquired. Note that in the current work, we only report results from the analysis of the behavioural data acquired during MR scanning. The visit ended with the participants receiving a reimbursement of $100 (see below for details).

Experimental design We developed a symmetric two-armed reversal bandit task, in which the available actions were not only associated with varying expected reward values but also with varying information gains (information-selective symmetric reversal bandit task, Figure 1a). More specifically, on each task trial participants could decide between the actions of choosing a square on the right of the computer screen vs. choosing a triangle on the left of the screen, or between the actions of choosing a square on the left vs. choosing a triangle on the right of the screen. Depending on the shape chosen, the action was either lucrative and returned a reward

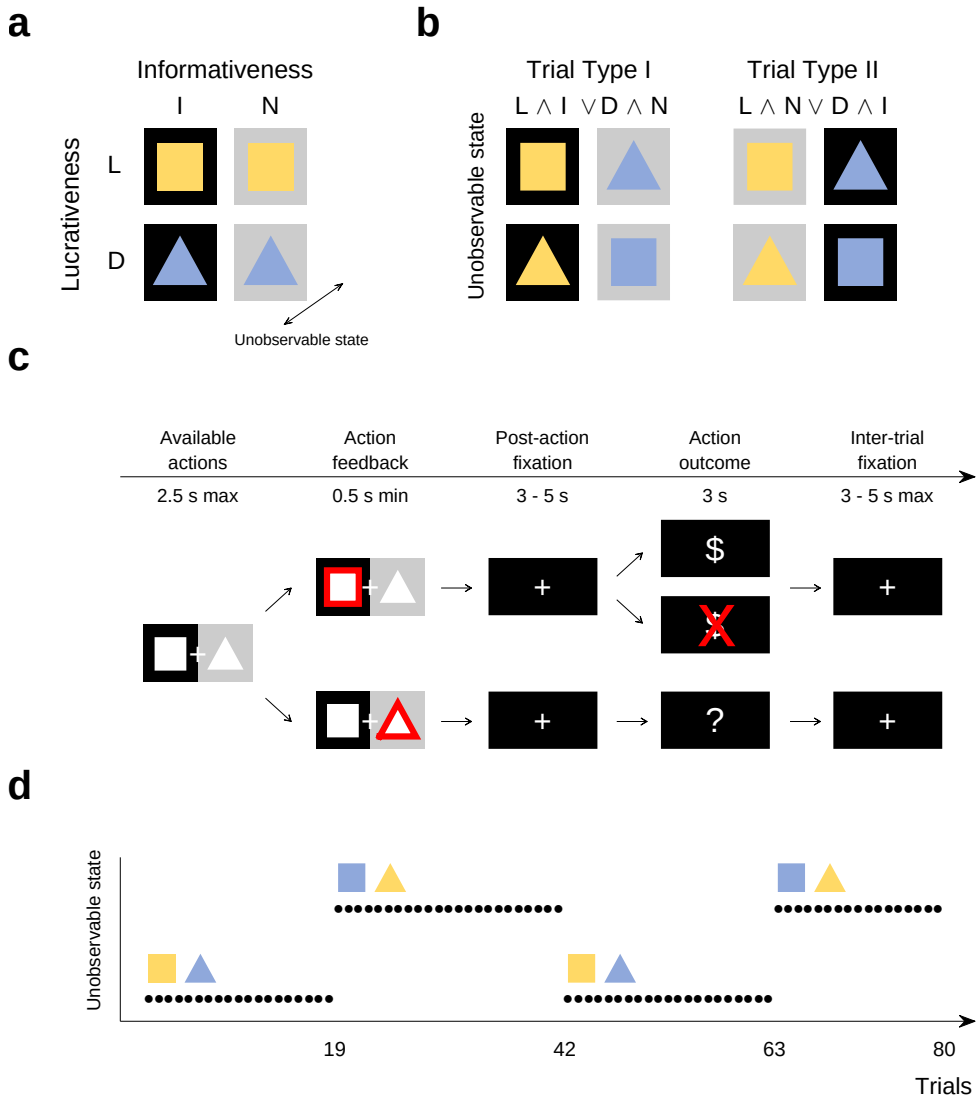

Figure 1. Information-selective symmetric reversal bandit task. **a** Experimental design. Possible actions differed in lucrativeness (lucrative (L) or detrimental (D)), as well as in informativeness (informative (I) or non-informative (N)). The former experimental factor was associated with selectable shapes (square and triangle), while the latter experimental factor was associated with black and grey screen sides. In addition, and unbeknownst to the participants, the lucrativeness of the shapes reversed at random trials throughout the experiment, corresponding to the unobservable task state: for some time, the square may represent the lucrative action, indicated here by its yellow color (and the triangle, accordingly, represent the detrimental action, indicated here by its blue color), but this could reverse. **b** On a given trial, participants faced a choice between either a lucrative and informative vs. a detrimental and non-informative action (Trial Type I; L ∧ I or D ∧ N) or a lucrative and non-informative vs. a detrimental and informative action (Trial Type II; L ∧ N or D ∧ I). **c** Trial design. Participants could indicate their choice within 2.5 seconds of the choice options onset. If they chose the shape on the black side, the returned reward was revealed (top). If they chose the shape on the grey side, the returned reward was not revealed (bottom). Note that the current lucrativeness of the shapes was not revealed to the participants, hence the white shape color. **d** Run design. Every 17 to 23 trials the reward probabilities associated with the shapes reversed. Here, the reversal times of the first run are shown. For trials 1 to 19 the square was lucrative, indicated by its yellow color, and the triangle was detrimental, indicated by its blue color. This reversed on trial 20 at which choosing the triangle became the lucrative action and choosing the square became the detrimental action. For the generation of this figure, please see *figure_ 1.m*.

of +1 with a probability of 0.85 and a reward of -1 with a probability of 0.15, or detrimental and returned a reward of +1 with a probability of 0.15 and a reward of -1 with a probability of 0.85. Depending on the side of the shape chosen, the action was also either informative and the returned reward was revealed to the participant, or it was non-informative and the returned reward was not revealed to the participant. Specifically, following an informative action either an image of a moneybag was displayed to signal a positive reward of +1, or an image of a crossed-out moneybag was displayed to signal the negative reward -1. In contrast, following a non-informative action an image of a question mark moneybag was displayed for both the rewards of +1 and -1. Importantly, while the actions' lucrativeness was not directly observable and could only be inferred from the revealed rewards, the actions' informativeness was directly observable throughout the experiment. In particular, for half of the participants the right screen side was associated with the informative action and the left screen side was associated with the non-informative action. For the other half of the participants the coupling between screen side and action informativeness was reversed. As a visual reminder for the participants, the informative and non-informative screen sides were also indicated by black and grey backgrounds, respectively. Note that we use the terms informative side and non-informative side in accordance with the action definitions. Similarly, we will also use the terms 'lucrative shape' and 'detrimental shape' instead of 'action of choosing the lucrative or detrimental shape' for simplicity. Also note that throughout the depiction of the experimental design and results, we visualize lucrative and detrimental actions by yellow and blue colors, respectively, and informative and non-informative actions by black and grey colors, respectively.

The experiment consisted of two runs of 80 trials each. On half of the trials, choosing the square was lucrative and choosing the triangle was detrimental. On the other half of the trials, choosing the square was detrimental and choosing the triangle was lucrative. We pseudo-randomized the sequence of lucrative shapes, such that choosing a certain shape was lucrative for 17-23 consecutive trials upon which the actions' lucrativeness reversed. This yielded a total of three shape lucrativeness reversals (or equivalently, four blocks of trials without a reversal) per task run (Figure 1d). Furthermore, we also pseudo-randomized the trial-by-trial sequence of choice options (e.g., a choice between the square on the informative side or the triangle on the non-informative side) with two constraints. First, a certain choice option combination occurred for a maximum of five consecutive trials. Second, on 50% of the trials in which the square was lucrative, the square was presented on the informative side (and the triangle on the non-informative side), while on the other 50% of the trials, the square was presented on the non-informative side (and the triangle on the informative side). The same constraint applied to those trials on which the triangle was lucrative. This way we did not only counterbalance the shape-side combinations, but also ensured that participants faced a choice between a lucrative and informative action (L $\wedge$ I) and a detrimental and non-informative (D $\wedge$ N) action on half of the trials. We refer to the trials with action choices between L $\wedge$ I and D $\wedge$ N actions as Trial Type I (Figure 1b). Accordingly, on the other half of the trials, participants faced a choice between a lucrative and non-informative action (L $\wedge$ N) and a detrimental and informative (D $\wedge$ I) action. We refer to the trials with action choices between L $\wedge$ N and D $\wedge$ I actions as Trial Type II (Figure 1b). Importantly, for a consistent experiment history across participants, we generated the sequence of lucrative shapes and choice options prior to the study and used the identical trial sequence for all participants. The task was implemented as *irb_ task.py* in Python 2.7 using PsychoPy V1.82.01 (Peirce, 2007).

Participants were encouraged to maximize the cumulative sum of returned rewards across all trials. As an incentive, participants were informed that in addition to a standard reimbursement of $70 for partaking in the study, they would receive a bonus up to $30 depending on their final

balance at the end of the second run of the task. They were not further informed about the balance-bonus conversion rate. In effect, however, all participants were payed the full bonus of $30 as requested by the Institutional Review Board.

Trial design Each trial started with the presentation of the two available choice options and participants were given a maximum of 2.5 seconds to indicate their choice (Figure 1c). If participants responded within this time window, the border of the chosen shape turned white to signal the recording of their choice. The duration of this feedback signal depended on the response time, such that the choice options and feedback together were presented for 3 seconds in total. Then, a post-choice fixation cross was presented for 3-5 seconds. This fixation cross was followed by the image representing the choice outcome, i.e., a moneybag, a crossed-out moneybag, or a question mark moneybag image, which was presented for 3 seconds. Finally, before the start of the next trial, an inter-trial fixation cross was displayed for 3-5 seconds. If participants did not respond within the choice time window, the message 'too slow' appeared for 0.5 seconds followed by an inter-trial fixation cross, a reward of -1 was automatically registered to the participant's account, and the next trial commenced. Notably, while the sequences of lucrative shapes and choice options were generated prior to the experiment, the fixation cross duration times and the returned rewards were sampled online as participants interacted with the task. Specifically, the fixation cross duration times were sampled uniformly from an interval of 3 to 5 seconds. The reward values +1 and -1 were sampled from discrete categorical distributions with probabilities 0.85 and 0.15 for the lucrative action and with probabilities 0.15 and 0.85 for the detrimental action, respectively.

Descriptive analyses

All data analyses were implemented in Matlab R2020a (The MathWorks, Natick, USA). We evaluated nine summary choice rates for every participant. In particular, we first evaluated overall and trial type-specific valid choice rates. These choice rates were defined as the number of valid action choices on all trials, on Type I trials, and on Type II trials divided by the number of all trials, of Type I trials, and of Type II trials, respectively. For example, by design there were 80 trials of Type I. If a participant missed to make a valid choice on one of these trials, the Trial Type I valid choice rate evaluated to 79/80. We then evaluated the choice rates for the lucrative and informative actions ($L \wedge I$), lucrative and non-informative actions ($L \wedge N$), detrimental and informative actions ($D \wedge I$), and detrimental and non-informative actions ($D \wedge N$). These choice rates were computed by dividing the number of respective actions by the number of valid choices of the corresponding trial type. Consequently, the choice rates of a given trial type are symmetrical, i.e., they sum up to 100%. For example, if a participant on Type I trials made 79 valid action choices of which 65 were $L \wedge I$ actions and 14 were $D \wedge N$ actions, then the $L \wedge I$ choice rate was 65/79 and the $D \wedge N$ choice rate was 14/79. In addition, we evaluated the choice rates of the lucrative actions and the informative actions. These were computed by dividing the sum of the number of $L \wedge I$ and $L \wedge N$ actions, as well as the sum of the number of $L \wedge I$ and $D \wedge I$ actions, by the number of valid choices on all trials. For example, if a participant made 159 valid choices in total, and of these 65 choices were $L \wedge I$ actions, while 58 were $L \wedge N$ actions, the lucrative action choice rate evaluated to 123/159. The nine individual summary choice rates were then averaged across participants to obtain the respective group summary choice rates and the standard error of the mean (SEM) over participants was evaluated.

In addition to the summary choice rates, we also evaluated trial-by-trial choice rates. Specifically, we computed group trial-by-trial $L \wedge I$, $L \wedge N$, $D \wedge I$, and $D \wedge N$ action choice rates.

To this end, for every trial we divided the number of respective actions by the number of valid choices on the trial over participants. As a given trial belonged to one of the two trial types, it either had associated L $\wedge$ I and D $\wedge$ N action choice rates, or associated L $\wedge$ N and D $\wedge$ I action choice rates. Consequently, in accordance with the summary action choice rates, the choice rates of each trial were symmetrical. For example, by design the first trial of the first run was of Type I for every participant. If on this trial 18 participants chose the L $\wedge$ I action, 5 chose the D $\wedge$ N action, and 1 participant missed to make a valid choice, the L $\wedge$ I action choice rate for this trial evaluated to 18/23 and the D $\wedge$ N action choice rate evaluated to 5/23. Finally, for each trial between two reversals, we computed the average reversal-locked group trial-by-trial L $\wedge$ I action and L $\wedge$ N action choice rates. Note that because the trial sequence was pseudo-randomized, the average reversal-locked group choice rate of a particular trial was computed based on different number of data points. For example, of the eight first trials, three were of Type I and had an associated group trial-by-trial L $\wedge$ I action choice rate, while five were of Type II and had an associated group trial-by-trial L $\wedge$ N action choice rate. Also note that as the number of trials between two reversals varied, there were fewer than eight 18th to 23rd trials.

Model formulation

Task model To render the task amenable to computational behavioural modelling, we first formulated a model of the task using concepts from the theory of partially observable Markov decision processes (Bertsekas, 2000). Specifically, we represent an experimental run by the tuple

$$M_{\text{Task}} := \left(T, S, A, R, O, p^{s_t^1, a_t}\left(r_t\right), f, g\right). \tag{1}$$

Here,

- T denotes the number of trials, indexed by $t = 1, ..., T$.
- $S := \mathbb{N}_2 \times \mathbb{N}_2$ denotes the set of states $s := \left(s^1, s^2\right)$. The first state component s^1 encodes the lucrative shape. Specifically, on trial t, s_t^1 takes on the value 1 if the square is lucrative and takes on the value 2 if the triangle is lucrative. From the perspective of the agent, s^1 is not directly observable. The second state component s^2 encodes the available actions. Specifically, on trial t, s_t^2 takes on the value 1, if the agent can choose between the square on the informative side or the triangle on the non-informative side. If on trial t the agent can choose between the square on the non-informative side or the triangle on the informative side, s_t^2 takes on the value 2. From the perspective of the agent, s^2 is directly observable.
- $A := \{A_1, A_2\}$ denotes the set of state-dependent action sets. Specifically, depending on the observable state component s_t^2 on a given trial t the available actions are either $A_1 := \{1, 4\}$ or $A_2 := \{2, 3\}$ for $s_t^2 = 1$ or $s_t^2 = 2$, respectively. If the available action set is A_1, then the agent can choose between $a = 1$, which corresponds to choosing the square on the informative side vs. $a = 4$, which corresponds to choosing the triangle on the non-informative side. If the available action set is A_2, then the agent can choose between $a = 2$, which corresponds to choosing the square on the non-informative side vs. $a = 3$, which corresponds to choosing the triangle on the informative side.
- $R := \{-1, +1\}$ denotes the set of rewards r.
- $O := \mathbb{N}_3$ denotes the set of observations o. $o = 1$ encodes the image of the crossed-out moneybag, $o = 2$ encodes the image of the moneybag, and $o = 3$ encodes the image of the question mark moneybag.

State- and action-dependent reward distribution

s_t^1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
a_t	1	1	2	2	3	3	4	4	1	1	2	2	3	3	4	4
r_t	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$
$p^{s_t^1,a_t}(r_t)$	0.15	0.85	0.15	0.85	0.85	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.15	0.85	0.15	0.85

Observation function

a_t	1	1	2	2	3	3	4	4
r_t	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$
$g(a_t, r_t)$	1	2	3	3	1	2	3	3

State-state transition distribution

s_t^1	1	1	2	2
s_{t+1}^1	1	2	1	2
$p\left(s_{t+1}^1 \mid s_t^1\right)$	0.9625	0.0375	0.0375	0.9625

Table 1. Model components. The upper table shows the state- and action-dependent reward distribution $p^{s_t^1,a_t}(r_t)$, the middle table shows the observation function g and the lower table shows the action-independent state-state transition distribution $p\left(s_{t+1}^1 \mid s_t^1\right)$.

- $p^{s_t^1,a_t}(r_t)$ is the state- and action-dependent reward distribution. For each combination of $s^1 \in S^1$ and $a \in A_{s^2}$, the state- and action-dependent reward distribution conforms to a discrete categorical distribution over r_t with probability parameters listed in the first panel of Table 1. As an example, consider $s^1 = 1$ (square is lucrative) and $a = 1$ (square on the informative side chosen). In this case, a reward of -1 is returned with a probability of 0.15 and a reward of $+1$ is returned with a probability of 0.85. On the other hand, if $s^1 = 2$ (triangle is lucrative) and $a = 1$ (square on the informative side chosen), the reward probabilities are reversed.

- f is the state evolution function, which specifies the value the state s_t takes on at trial t,

$$f : \mathbb{N}_T \to S, t \mapsto f(t) := s_t. \tag{2}$$

  f is defined in a tabular form and corresponds to the sequence of lucrative shapes and choice options presented to all participants (cf. Section S.3: Experimental state sequence).

- g is the observation function

$$g : A \times R \to O, (a, r) \mapsto g(a, r) := o \tag{3}$$

  as defined in the second panel of Table 1. For the informative actions $a = 1$ and $a = 3$, g is injective: The reward $r = -1$ is mapped onto the observation $o = 1$, corresponding to the image of the crossed-out moneybag, while the reward $r = +1$ is mapped onto the observation $o = 2$, corresponding to the image of the moneybag. For the non-informative actions $a = 2$ and $a = 4$, g is not injective: both rewards $r = -1$ and $r = +1$ are mapped onto the observation $o = 3$, corresponding to the image of the question mark moneybag.

Agent models We designed five agent models, denoted by C1, C2, A1, A2, and A3, to account for the putative cognitive processes underlying participants' choices. Before we introduce the individual characteristics of these agents, we first represent the general structure of an agent

interacting with an experimental run. This general agent structure corresponds to the tuple

$$M_{\text{Agent}} := \left(T, S, A, R, O, p\left(s_1^1\right), p\left(s_{t+1}^1 | s_t^1\right), p^{a_t}\left(o_t | s_t^1\right), p^{a_t}\left(r_t | s_t^1\right)\right). \tag{4}$$

Here,

- T, S, A, R and O are defined as the corresponding sets of the task model M_{Task}.
- $p\left(s_1^1\right)$ denotes the initial agent belief state, which specifies the agent's subjective uncertainty over the non-observable state component s_1^1 at trial $t = 1$. $p\left(s_1^1\right)$ is defined in terms of the discrete categorical distribution

$$p(s_1^1 = 1) = 0.5 \text{ and } p(s_1^1 = 2) = 0.5. \tag{5}$$

  Because $p(s_1^1)$ is fully parameterized by specifying $p(s_1^1 = 1)$, we hereinafter occasionally represent the initial belief state by the scalar $b_1 := p(s_1^1 = 1)$.
- $p\left(s_{t+1}^1 | s_t^1\right)$ is the state-state transition distribution, which specifies the agent's subjective uncertainty over the non-observable state component s_{t+1}^1 at trial $t+1$ given the non-observable state component s^1 at trial t. More specifically, for each $s^1 \in S^1$, the state-state transition distribution corresponds to a discrete categorical distribution over s_{t+1}^1 with probability parameters listed in the third panel of Table 1. Note that the trial-by-trial state transitions are probabilistic, because from the agent's perspective a reversal in the shapes' lucrativeness could happen between any two trials. This is in contrast with the state evolution from the task perspective, which - given the apriori defined sequence of lucrative shapes - is deterministic (cf. eq. (2)). Crucially, participants were informed that a reversal would happen 1-4 times in a run, but were not informed about the approximate number of trials without a reversal. Therefore, we equipped the agent with a constant reversal probability of 0.0375, which reflects the true reversal frequency in a run (there were 3 reversals across the 80 trials). For example, if $s_t^1 = 1$ (square is lucrative), the agent allocates a probability of 0.9625 to the event that on the next trial s_{t+1}^1 takes on the value 1 (square is lucrative), while it allocates a probability of 0.0375 to the event that s_{t+1}^1 takes on the value 2 (triangle is lucrative).
- $p^{a_t}\left(r_t | s_t^1\right)$ is the action-dependent state-conditional reward distribution, which specifies the agent's subjective uncertainty over the reward r_t given the non-observable state component s^1 and action a at trial t. More specifically, for each combination of $s^1 \in S^1$ and $a \in A_{s^2}$, the action-dependent state-conditional reward distribution defines a discrete categorical distribution over r_t with probability parameters corresponding to

$$p^{a_t=a}\left(r_t = r | s_t^1 = s^1\right) := p^{s_t^1=s^1, a_t=a}\left(r_t = r\right). \tag{6}$$

  Notice that the only difference between the agent's action-dependent state-conditional reward distribution and the task's state- and action-dependent reward distribution is that for the former, the state is conceived as a random variable, while for the latter the state is conceived as a parameter. We equipped the agent with the true reward emission probabilities to reflect the task instructions. In particular, participants were truthfully informed that choosing the lucrative shape would return a reward of +1 with a high probability and a reward of -1 with a low probability, as well as that choosing the detrimental shape would return a reward of +1 with a low probability and a reward of -1 with a high probability.
- $p^{a_t}\left(o_t | s_t^1\right)$ is the action-dependent state-conditional observation distribution, which specifies

the agent's subjective uncertainty over the observation o_t given the non-observable state component s^1 and action a at trial t. In detail, for each combination of $s^1 \in S^1$ and $a \in A_{s^2}$, the action-dependent state-conditional observation distribution corresponds to a discrete categorical distribution over o_t with probability parameters resulting from transforming the distribution of r_t by the observation function g. Formally,

$$p^{a_t=a}\left(o_t = o | s_t^1 = s^1\right) := \sum_{\{r|g(a,r)=o\}} p^{a_t=a}\left(r_t = r | s_t^1 = s^1\right). \tag{7}$$

For the informative actions $a \in \{1, 3\}$, it thus follows that

$$p^{a_t=a}\left(o_t = 1 | s_t^1 = s^1\right) = p^{a_t=a}\left(r_t = -1 | s_t^1 = s^1\right) \tag{8}$$

and

$$p^{a_t=a}\left(o_t = 2 | s_t^1 = s^1\right) = p^{a_t=a}\left(r_t = +1 | s_t^1 = s^1\right). \tag{9}$$

For the non-informative actions $a \in \{2, 4\}$, on the other hand, it follows that

$$p^{a_t=a}\left(o_t = 3 | s_t^1 = s^1\right) = p^{a_t=a}\left(r_t = -1 | s_t^1 = s^1\right) + p^{a_t=a}\left(r_t = 1 | s_t^1 = s^1\right) = 1. \tag{10}$$

As an example, consider the case $s^1 = 1$ (square is lucrative) and $a = 1$ (square on the informative side chosen). The agent allocates the same probabilities to observing either the image of the crossed-out moneybag or the image of the moneybag as to obtaining a reward of -1 or +1, respectively. Alternatively, consider the case $s^1 = 1$ (square is lucrative) and $a = 4$ (triangle on the non-informative side chosen). In this case, the agent allocates a probability of 1 to observing the image of the question mark moneybag.

Based on the general agent structure encoded in the tuple M_{Agent}, we next discuss our model space of interest, which comprises two control agents, denoted by C1 and C2, and three Bayesian agents, denoted by A1, A2, and A3.

Control agents C1 and C2 The control agents C1 and C2 rely on heuristic choice strategies. Because C1 and C2 do not represent a belief state, their action valence function is a function of action only,

$$v : A \to \mathbb{R}, a \mapsto v(a). \tag{11}$$

To realize an action on trial t, both agents use a probabilistic decision rule. Specifically, C1 and C2 directly translate the action valences into action and observation history-dependent choice probabilities.

C1: A belief state-free random choice agent Agent C1 may be considered a cognitive null model. It does not have an optimization aim based on which it could differentiate between actions, but merely allocates equal valences to all available actions $a \in A_{s^2}$,

$$v_{\text{C1}}(a) := \frac{1}{|A_{s^2}|} = 0.5. \tag{12}$$

C2: A belief state-free win-stay-lose-switch agent Agent C2 aims to maximize immediate rewards without relying on a belief state. To this end, C2 adopts a heuristic win-stay-lose-switch strategy (Robbins, 1952). Specifically, on each trial t, agent C2 determines its preferred choice based on previous reward signaling observations, but does not take action informativeness (i.e.,

shape laterality) into account. Formally, on trial $t = 1$, the action valence function of agent C2 is defined as

$$v^1_{\text{C2}}(a) := 0.5 \text{ for all } a \in A_{s^2_1}. \tag{13}$$

Subsequently, on trials $t = 2, 3, ..., T$ agent C2 allocates action valences according to

$$v^t_{\text{C2}}(a) := \begin{cases} 0, & \text{if } o_{t-1} = 1 \text{ and } a \in \mathcal{A}_{t-1} \text{ or } o_{t-1} = 2 \text{ and } a \notin \mathcal{A}_{t-1}, \\ 1, & \text{if } o_{t-1} = 2 \text{ and } a \in \mathcal{A}_{t-1} \text{ or } o_{t-1} = 1 \text{ and } a \notin \mathcal{A}_{t-1}, \\ v^{t-1}_{\text{C2}}(a), & \text{if } o_{t-1} = 3, \end{cases} \tag{14}$$

where $\mathcal{A}$ denotes the set of actions of choosing a given shape and thus $\mathcal{A} := \{1, 2\}$ for the actions choose square and $\mathcal{A} := \{3, 4\}$ for the actions choose triangle.

Informally, agent C2 allocates equal initial action valences to all actions, because on trial $t = 1$ no previous observations are available and therefore the agent has no basis for differentiating between actions. Subsequently, on trials $t = 2, 3, ..., T$, agent C2 allocates action valences depending on the observation on trial $t - 1$. Specifically, if on trial $t - 1$ the choice of a shape resulted in the observation $o = 1$, i.e., the image of the crossed-out moneybag, then agent C2 allocates an action valence of 0 to choosing the same shape and an action valence of 1 to choosing the other shape on trial t. In contrast, if on trial $t - 1$ the choice of a shape resulted in the observation $o = 2$, i.e., the image of the moneybag, then agent C2 allocates an action valence of 1 to choosing the same shape and an action valence of 0 to choosing the other shape on trial t. Crucially, if on trial $t - 1$ the choice of a shape resulted in the observation $o = 3$, i.e., the image of the question mark moneybag, the value of the returned reward is not signaled to the agent. In this case, agent C2 relies on its action valence allocation scheme from trial $t - 1$: The action valences that agent C2 allocates to choosing a given shape on trial t correspond to the valences the agent allocated to choosing that shape on trial $t - 1$.

Bayesian agents (A1, A2 and A3) The Bayesian agents maintain a belief state which subserves their action choice. Specifically, the distributions $p\left(s^1_1\right)$, $p\left(s^1_{t+1}|s^1_t\right)$ and $p^{a_t}\left(o_t|s^1_t\right)$ of M_{Agent} induce a trial-by-trial action-dependent joint probability distribution $p^{a_{1:t-1}}\left(s^1_{1:t}, o_{1:t-1}\right)$. This allows for the recursive evaluation of the belief state $p^{a_{1:t-1}}\left(s^1_t|o_{1:t-1}\right)$ on trial t given the history of observations $o_{1:t-1}$ and actions $a_{1:t-1}$ by means of

$$p^{a_{1:t-1}}\left(s^1_t|o_{1:t-1}\right) = \frac{\sum_{s^1_{t-1}} p\left(s^1_t|s^1_{t-1}\right) p^{a_{t-1}}\left(o_{t-1}|s^1_{t-1}\right) p^{a_{1:t-2}}\left(s^1_{t-1}|o_{1:t-2}\right)}{\sum_{s^1_t}\sum_{s^1_{t-1}} p\left(s^1_t|s^1_{t-1}\right) p^{a_{t-1}}\left(o_{t-1}|s^1_{t-1}\right) p^{a_{1:t-2}}\left(s^1_{t-1}|o_{1:t-2}\right)}, \tag{15}$$

for $t = 1, ..., T$, and where the prior belief state is given by $p\left(s^1_1\right)$ for $t = 1$. For the derivation of eq. (15), please see Section S.4: Belief state, posterior predictive distribution, and KL-divergence. Intuitively, the Bayesian agents update their belief state in a trial-by-trial fashion based on the observation made after choosing a shape on either side and by accounting for a reversal in the shapes' lucrativeness. Note that on an implementational level, we represent the distributions $p\left(s^1_1\right)$, $p\left(s^1_{t+1}|s^1_t\right)$ and $p^{a_t}\left(o_t|s^1_t\right)$ by stochastic matrices and evaluate the belief state using matrix multiplication as detailed in Section S.5: Belief state and posterior predictive distribution implementation.

Based on their belief state representation, the Bayesian agents then decide for an action based on a combination of an action valence function, which evaluates the desirability of a given action in the light of the agent's current belief state, and a decision function, which selects the maximal

desirable action as the action to issue. Specifically, the scalar representation of the belief state

$$b_t := p^{a_{1:t-1}}\left(s_t^1 = 1 | o_{1:t-1}\right) \tag{16}$$

constitutes the basis for action evaluation by means of an action valence function

$$v : A \times [0,1] \to \mathbb{R}, (a,b) \mapsto v\left(a,b\right). \tag{17}$$

As detailed below, the exact forms of the valence function differ between agents A1, A2, and A3. However, to realize an action, all Bayesian agents pass the evaluated action valences on to a maximizing decision rule of the form

$$d : \mathbb{R} \times [0,1] \to A_{s^2}, v(\cdot, b) \mapsto d(v(\cdot, b)) := \underset{a \in A_{s^2}}{\arg\max}\, v(a,b). \tag{18}$$

On every trial, the Bayesian agents thus choose the action with the highest valence.

A1: A belief state-based exploitative agent Agent A1 uses its belief state to maximize the immediate reward gain. To this end, agent A1 uses an action valence function that allocates to action $a_t = a$ an action valence based on the action-dependent expected reward under the current belief state $b_t = b$ according to

$$v_{\mathrm{A1}}\left(a,b\right) := b\mathbb{E}_{p^a\left(r_t|s_t^1=1\right)}\left(r_t\right) + \left(1-b\right)\mathbb{E}_{p^a\left(r_t|s_t^1=2\right)}\left(r_t\right). \tag{19}$$

The upper and lower panels of Figure 2a visualize the A1 valences for actions $a \in A_1$ (choose square on the informative side or triangle on the non-informative sidel) and $a \in A_2$ (choose square on the non-informative side or triangle on the informative side), respectively, as functions of the belief state b. Note that the expected reward is

$$\mathbb{E}_{p^a(r_t|s_t)}\left(r_t\right) = 0.85 \cdot -1 + 0.15 \cdot 1 = -0.7 \tag{20}$$

for choosing the detrimental shape and

$$\mathbb{E}_{p^a(r_t|s_t)}\left(r_t\right) = 0.85 \cdot 1 + 0.15 \cdot -1 = 0.7 \tag{21}$$

for choosing the lucrative shape. Consequently, the more certain A1 becomes that a given shape is lucrative (as b gets closer to 0 or 1 from 0.5) the higher the belief state-weighted expected reward for choosing that shape and, accordingly, the lower the belief state-weighted expected reward for choosing the other shape. As the belief state-weighted expected reward is irrespective of the side of the shape, in the case of both sets of available actions, agent A1 allocates valences without taking the actions' informativeness into account.

A2: A belief state-based explorative agent Agent A2 explores its belief state to maximize the immediate information gain. To this end, on trial t agent A2 allocates a valence to each available action $a_t = a$ based on its expected Bayesian surprise

$$v_{\mathrm{A2}}\left(a,b\right) := \sum_{o_t} p^{a_{1:t-1},a_t=a}\left(o_t|o_{1:t-1}\right) \mathrm{KL}\left(p^{a_{1:t-1},a_t=a}\left(s_{t+1}^1|o_{1:t-1},o_t\right)\middle|\middle|p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right)\right). \tag{22}$$

The first term in eq. (22),

$$p^{a_{1:t-1},a}\left(o_t|o_{1:t-1}\right) = bp^a\left(o_t|s_t^1=1\right) + (1-b)\,p^a\left(o_t|s_t^1=2\right) \tag{23}$$

denotes the agent's belief state-dependent posterior predictive distribution, which specifies the agent's subjective uncertainty over the observation o_t given action a on trial t and its history of observations $o_{1:t-1}$ and actions $a_{1:t-1}$. For a derivation of the right-hand side of eq. (23), please see Section S.4: Belief state, posterior predictive distribution, and KL-divergence and for implementational details regarding its evaluation, please see Section S.5: Belief state and posterior predictive distribution implementation. The second term in eq. (22),

$$\mathrm{KL}\left(p^{a_{1:t-1},a_t=a}\left(s_{t+1}^1|o_{1:t-1},o_t\right)\middle|\middle|p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right)\right) = (1-b^{a,o})\ln\left(\frac{1-b^{a,o}}{1-b}\right) + b^{a,o}\ln\left(\frac{b^{a,o}}{b}\right) \tag{24}$$

denotes the Kullback-Leibler (KL) divergence between the agent's belief state on trial t and its virtual belief state on trial $t+1$ given observation o_t and action $a_t = a$ on trial t. Intuitively, this KL divergence quantifies the information gain afforded by choosing action $a_t = a$ and making observation o_t on trial t. On the right-hand side of eq. (24), b denotes the agent's belief state on trial t (cf. eq. (19)) and $b^{a,o}$ denotes the agent's belief state b_{t+1} resulting from action $a_t = a$ and observation $o_t = o$. For a derivation of the right-hand side of eq. (24), please see Section S.4: Belief state, posterior predictive distribution, and KL-divergence. In summary, $v_{\mathrm{A2}}(a,b)$ quantifies the agent's weighted expected information gain afforded by choosing $a_t = a$ given its current belief state $b_t = b$, where the expectation is formed with regards to the possible observations o_t on trial t and the weighting is determined by the agent's current estimate of observing a specific observation $o_t = o$.

The upper and lower panels of Figure 2b visualize the A2 valences for actions $a \in A_1$ (choose square on the informative side or triangle on the non-informative side) and $a \in A_2$ (choose square on the non-informative side or triangle on the informative side), respectively, as functions of the belief state b. Choosing the shape on the non-informative side does not deliver reward information. Therefore, the expected Bayesian surprise-based A2 valence is always higher for the informative action, irrespective of the agent's belief state. Yet, the difference between the informative and non-informative action valences depends on the belief state. Specifically, in contrast to agent A1, the more uncertain agent A2 becomes about the lucrative shape (as b gets closer to 0.5 from 1 or 0), the larger the difference between the valences and thus the stronger the agent's preference for the informative action.

A3: A belief state-based explorative-exploitative hybrid agent Agent A3 combines the choice strategies of agents A1 and A2 and uses its belief state to maximize the combination of immediate reward gain and information gain. Formally, on each trial t and for each available action $a_t = a \in A_{s^2}$, agent A3 evaluates its action valences based on the convex combination of the action valences of agents A1 and A2,

$$v_{\mathrm{A3}}(a,b) := \lambda v_{\mathrm{A1}}(a,b) + (1-\lambda)\,v_{\mathrm{A2}}(a,b), \tag{25}$$

where $\lambda \in [0,1]$ is the weighting parameter. The uppwer and lower panels of Figures Figure 2c and Figure 2d visualize valences of agent A3 for actions $a \in A_1$ (choose square on the informative side or triangle on the non-informative side) and $a \in A_2$ (choose square on the non-informative side or triangle on the informative side), respectively, as functions of the belief state b for weight

parameter values of $\lambda = 0.5$ and $\lambda = 0.25$, respectively. For $\lambda = 1$, the action valences of agent A3 correspond to the action valences of A1, while for $\lambda = 0$ the action valences of agent A3 correspond to the action valences of agent A2. For weighting parameter values $\lambda \in]0, 1[$, the decision strategy of agent A3 results from a mixture of the strategies of agents A1 and A2: For non-extreme belief state values, i.e., b values close to 0.5, agent A3 allocates a higher valence to choosing the shape on the informative side, even if the agent allocates a lower probability to that shape being lucrative. This shows the contribution of agent A2's choice strategy. For more extreme belief state values, i.e., b values close to 0 or 1, agent A3 allocates a higher valence to choosing the shape with the higher probability to be lucrative, even if the action is non-informative. This shows the contribution of agent A1's choice strategy. Note, however, that a λ value of 0.5 should not be understood as the agent A3's choice strategy resembling respective strategies of agents A1 of A2 to equal degree. The reason for this is that agent A3 applies a convex combination of the A1 and A2 action valences, which take values in different ranges (-0.7 to 0.7 for A1, 0 to 0.23 for A2). Therefore, while for $\lambda = 0.5$ the action valences of agent A3 primarily reflect the contribution of agent A1 (cf. Figure 2c), the contribution of the action valences of agent A2 becomes evident for $\lambda = 0.25$ (cf. Figure 2d).

Model evaluation

Data analysis models To evaluate the agent models in light of the participants' data, we embedded the agent models in a statistical inference framework. In particular, for agent models C2, A1, A2, and A3 we formulated behavioural data analysis models by nesting the agent-specific action valence functions in a softmax operation (Reverdy & Leonard, 2015). To this end, we defined the probability of action a given the history of actions $a_{1:t-1}$ and observations $o_{1:t-1}$ as

$$p\left(a_t = a | a_{1:t-1}, o_{1:t-1}\right) := \frac{\exp\left(\tau^{-1} v(a, \cdot)\right)}{\sum_{\tilde{a} \in A_{s^2}} \exp\left(\tau^{-1} v(\tilde{a}, \cdot)\right)}, \tag{26}$$

where for each agent, $v(a, \cdot)$ corresponds to the agent-specific action valence function. Here, the parameter $\tau \in \mathbb{R}_{>0}$ encodes the level of post-decision noise: The lower the value of τ, the higher the probability that the action with the higher action valence is realized and thus the lower the post-decision noise. Notably, as agent C1 allocates equal action valences throughout, for any τ value the softmax operation returns a uniform probability distribution. Therefore, for this agent a softmax operation is not required, and we defined the respective data analysis model as

$$p\left(a_t = a | a_{1:t-1}, o_{1:t-1}\right) = p\left(a_t = a | a_{1:t-1}\right) := v_{\mathrm{C1}}\left(a\right). \tag{27}$$

Model estimation To estimate the parameters of the data analysis models for agents C2, A1, A2, and A3, we used a maximum likelihood (ML) approach. Specifically, we assumed conditionally independently distributed actions and for each participant's data defined the conditional log likelihood function as

$$\ell : \Theta \to \mathbb{R}, \theta \mapsto \ell(\theta) := \ln \prod_{n=1}^{N} p^{\theta}\left(a_n | a_{1:n-1}, o_{1:n-1}\right) = \sum_{n=1}^{N} \ln p^{\theta}\left(a_n | a_{1:n-1}, o_{1:n-1}\right). \tag{28}$$

For agents C2, A1, and A2, the conditional log likelihood function is a function of the softmax operation parameter τ only. Thus, for these data analysis models $\theta := \tau$. For agent A3, the conditional log likelihood function is a function of τ and the action valence weighting parameter $\lambda \in [0, 1]$ (cf. eq. (25)). Thus, for this data analysis model $\theta := (\tau, \lambda)$. Note that in eq. (28) we

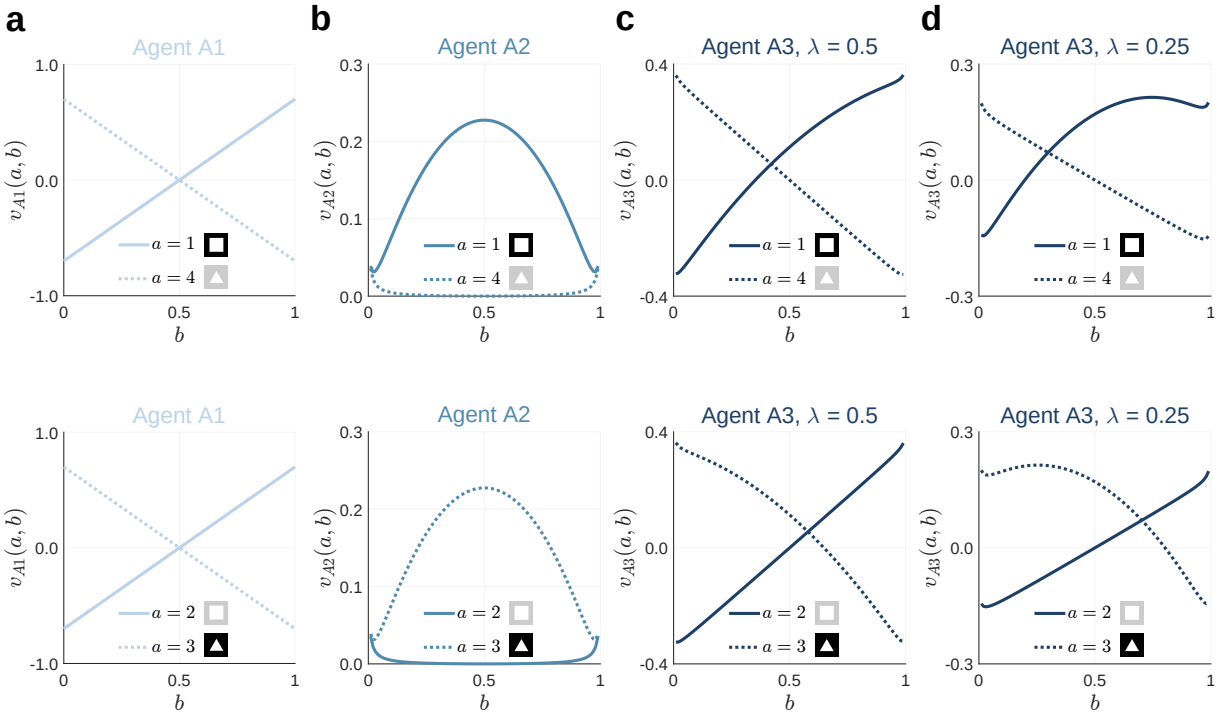

Figure 2. Bayesian agents' action valence functions. **a** Action valence functions of agent A1 for the available action set $A_1 := \{1, 4\}$ (upper panel) and the available action set $A_2 := \{2, 3\}$ (lower panel). Agent A1 allocates action valences based on the belief state-weighted expected reward. As the expected rewards for choosing the lucrative or detrimental shape are constant, the more extreme the agent's belief that a given shape is lucrative the higher the valence it allocates to choosing the corresponding shape and the lower the valence it allocates to choosing the other shape. The valences of A1 do not depend on the actions' informativeness, which reverses between available action sets A_1 and A_2 and therefore the two panels are identical. **b** Action valence functions of agent A2 for the available action set $A_1 := \{1, 4\}$ (upper panel) and the available action set $A_2 := \{2, 3\}$ (lower panel). Agent A2 allocates action valences based on the expected Bayesian surprise, which is higher for the informative action than for the non-informative action, thus the graphs on the upper and lower panels flip. The lower the agent's uncertainty about the lucrative shape, the larger the benefit of the informative action. **c, d** Action valence functions of agent A3 for the available action set $A_1 := \{1, 4\}$ (upper panel) and the available action set $A_2 := \{2, 3\}$ (lower panel) for $\lambda = 0.5$ (**c**) and $\lambda = 0.25$ **d** Agent A3 allocates action valences based on the convex combination A1 and A2 action valences. The higher the value of λ the more the valences of A3 resemble the valences of A1 and correspondingly, the lower the value of λ the more the valences of A3 resemble the valences of A2. Note that the colors used for the graphical depiction of the agents' action valence functions correspond to the agent model color scheme used for model recovery and model comparison in all Figures below. For implementational details, please see *figure_ 2.m*.

index trials by $n = 1, ..., N$ rather than by $t = 1, ..., 2T$ to account for the fact that participant's occasionally failed to respond, resulting in invalid choice trials. n hence indexes a participant's nth valid choice trial and N denotes a participant's total number of valid choice trials.

For every data analysis model, we estimated θ based on single participant data sets by minimizing the respective negative conditional log likelihood function using Matlab's constrained nonlinear optimization function *fmincon.m* (Byrd et al., 1999, 2000; Waltz et al., 2006). The parameter constraints were set to $\tau \in [0.01, 2.5]$ and to $\lambda \in [0, 1]$. To mitigate the risk of identifying local maximizers rather than global maximizers of the conditional log likelihood function, the initial parameter estimate values were sampled from a continuous uniform distribution covering the constrained parameter space, the parameter estimation procedure was repeated ten times, and only the parameter estimates achieving the highest conditional log likelihood function value were regarded as ML estimates (Wilson & Collins, 2019). Note that the conditional log likelihood

function of agent model C1 does not require any parameter optimization and can be evaluated directly.

Model comparison To compare the models' relative plausibilities given the participant's data, we first evaluated all agent model- and participant-specific Bayesian Information Criterion scores according to (Schwarz et al., 1978)

$$\text{BIC} := \ell(\hat{\theta}) - \frac{k}{2} \ln N. \tag{29}$$

Here, $\hat{\theta}$ denotes the ML parameter estimate, k denotes the respective model's number of to be estimated parameters, and N denotes the respective participant's total number of valid choices. The BIC scores of all agents and participants were then subjected to a random-effects Bayesian model selection procedure as implemented in *spm_ BMS.m* and distributed with the SPM toolbox for neuroimaging data analysis (`www.fil.ion.ucl.ac.uk/spm/`, Stephan et al. (2009); Rigoux et al. (2014)). For each model, *spm_ BMS.m* returns a protected exceedance probability (PEP) which corresponds to the group-level probability that the particular model is more likely than the other models of the model space.

Model and parameter recovery analyses

Model and parameter recovery To validate our agent-based behavioural modelling approach, we performed a number of model recovery simulations with the aim of assessing to which degree our analysis approach allows for reliably arbitrating between the models in our model space. To this end, we first generated synthetic behavioural data using each of our agent-based data analysis models C1, C2, A1, A2, and A3. The synthetic behavioural data sets comprised agent-specific actions on 160 trials under the experimentally employed state sequence (cf. Section S.3) and for agent action-dependent observations $o_{1:2T}$ and rewards $r_{1:2T}$. For agent C1, the synthetic data were generated without further specifications. For agents C2, A1, A2, and A3, the synthetic data were generated using softmax operation parameter values between $\tau = 0.05$ and $\tau = 2.5$ in steps of $\Delta\tau = 0.05$. In addition, for agent A3, the synthetic data were generated with parameter values $\lambda \in \{0.1, 0.25, 0.3, 0.5, 0.7, 0.9\}$. For each data generating model and for each parameter value, we generated 24 data sets (corresponding to the experimental sample size), evaluated these data sets using all models in our model space by means of the ML estimation and BIC model comparison approach discussed above, and repeated this procedure 10 times. Finally, as our primary outcome measure, we evaluated the average protected exceedance probabilities across the 10 data generation and data analysis repeats. Additionally, to assess to which degree we can reliably estimate the data analysis model parameters, we performed a number of parameter recovery analyses. Specifically, for a given data generating model and model parameter setting, we evaluated the average ML parameter estimates and their associated standard errors of the data generating model across the 24 synthetic data sets and the 10 data generation and data analysis repeats.

Post-hoc model validation Upon identifying agent A3 as the most plausible model to explain our observed experimental data based on the formal criteria discussed above, we further investigated the behavioural face validity and the participant parameter estimate-based identifiability of this model. To this end, we first generated 24 synthetic data sets based on agent A3 using the 24 participant-specific ML parameter estimates. As in the model recovery analysis, the synthetic data conformed to the agent's actions on 160 trials, with trial sequence identical

to the participants' trial sequence and with agent action-dependent observations. To assess the model's behavioural face validity, we then subjected these 24 data sets to the descriptive analyses described above and computed the same action choice rates as for the participants' data. To minimize random variation, we repeated this process 100 times and for each action choice rate evaluated the average over the 100 simulations. Similarly, we subjected the simulated data sets to our model estimation and comparison procedures and evaluated the average model and parameter recoverability performance as discussed above.

Results

In the following, we first report a set of summary statistics of the participants' choice behavior. We then discuss the model and parameter recovery study results. Finally, we report the results of our model-based experimental data analysis and its post-hoc validation.

Descriptive statistics We first evaluated a set summary action choice rates and their associated SEMs. Participants completed the vast majority of trials and achieved an overall valid choice rate of $97.86\% \pm 0.62$. There was virtually no difference in the number of valid choices between trial types: The valid choice rate on Trial Type I was $97.97\% \pm 0.62$ and the valid choice rate on Trial Type II was $97.76\% \pm 0.68$. As shown in Figure 3a, on Trial Type I, the majority of action choices was lucrative and informative ($L \wedge I$, $87.45\% \pm 1.53$), while only a few action choices were detrimental and non-informative ($D \wedge N$, $12.55\% \pm 1.53$). The difference between the choice rates on Trial Type II was less pronounced: as shown in Figure 3a, $66.01\% \pm 2.28$ of the action choices on Trial Type II were lucrative and non-informative ($L \wedge N$), while $33.99\% \pm 2.28$ were detrimental and informative ($D \wedge I$). Summed over informative and non-informative action choices, the lucrative action choice rate was $76.74\% \pm 1.7$, whereas summed over lucrative and detrimental action choices, the informative action choice rate was $60.74\% \pm 0.92$. Notably, participants made more lucrative choices if the lucrative action was also informative ($L \wedge I$, $87.45\% \pm 1.53$) compared to lucrative choices if the lucrative action was non-informative ($L \wedge N$, $12.55\% \pm 1.53$). To statistically corroborate this finding, we conducted a two-sided paired sample t-test across participants. This yielded a test statistic of $t(23) = 11.55$ with an associated p-value smaller than 0.001. Taken together, these summary statistics suggest that while participants' choices were primarily guided by action lucrativeness, participants also took the action's informativeness into account when deliberating which action to choose.

We next evaluated group trial-by-trial choice rates. As show in Figure 3b, on the majority of trials, the two lucrative action choice rates $L \wedge I$ and $L \wedge N$ prevailed over the two detrimental action choice rates $D \wedge I$ and $D \wedge N$. This effect was more pronounced for the trial-by-trial $L \wedge I$ action choice rate. Notably, as shown in Figure 3c, both the average $L \wedge I$ action choice rate as well as the average $L \wedge N$ action choice rate exhibit an overall increase between two reward rate reversals, indicating that participants were able to gradually resolve their uncertainty about the currently lucrative shape. Moreover, as shown in Figure 3d, although the average $L \wedge I$ action choice rate was larger than the average $L \wedge N$ action choice rate on virtually all trials between two reversals, their difference decreased slightly between the first trial after and the last trial before a reward rate reversal. This suggests that with decreasing uncertainty about the currently lucrative shape participants placed less valence on the actions' informativeness.

Model recovery analyses We validated our modelling initiative by conducting a series recovery analyses of the agent-based data analysis models and their parameters. Figure 4 summarizes

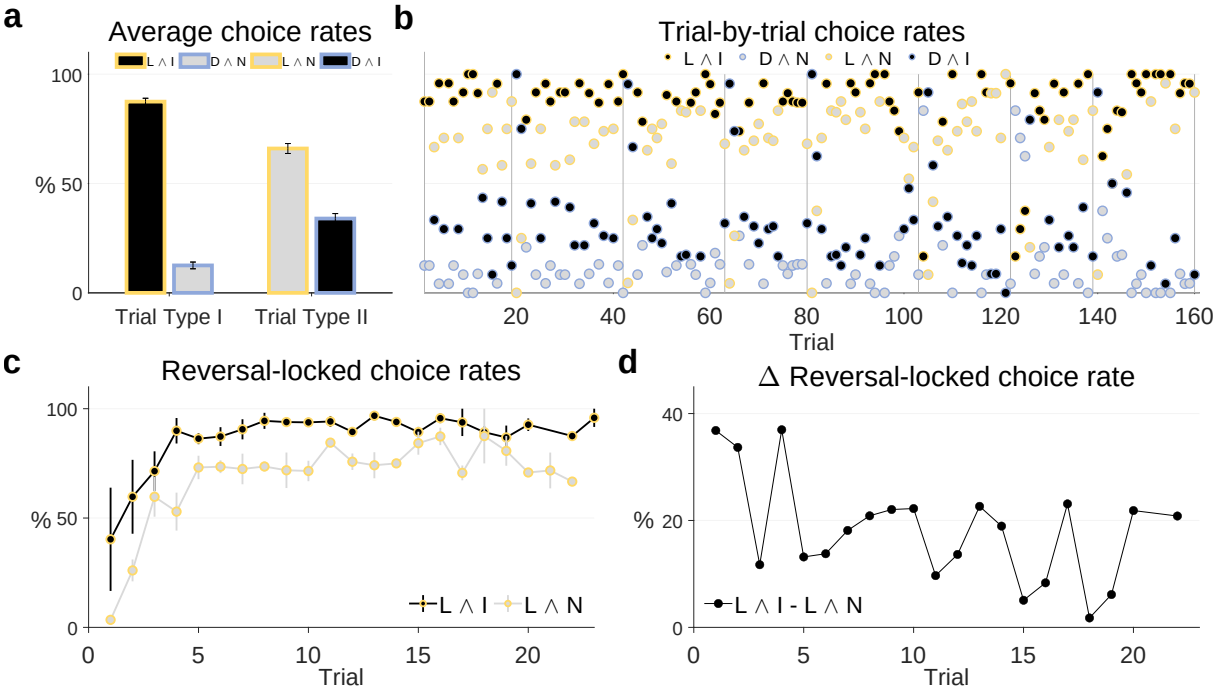

Figure 3. Participant choice rates. **a** Average choice rates across participants and task dynamics for Trial Type I and Trial Type II. On Trial Type I, participants faced a choice between a lucrative and informative (L ∧ I) and a detrimental and non-informative (D ∧ N) action. On average, participants preferred the L ∧ I action. On Trial Type II, participants faced a choice between a lucrative and non-informative (L ∧ N) and a detrimental and informative (D ∧ I) action. On average, participants preferred the L ∧ N action, but to a lesser degree than the L ∧ I action on Trial Type I. **b** Group trial-by-trial L ∧ I, D ∧ N, L ∧ N, and D ∧ I action choice rates. The vertical lines represent the last trial before a reward rate reversal. The vertical lines at $t = 80$ and $t = 160$ mark the end of the first and second run, respectively. Note that given the experimental design, the action choice rates on Trial Type I (L ∧ I vs. D ∧ N) and on Trial Type II (L ∧ N vs. D ∧ I) are complementary and add up to 100%. On the majority of trials the lucrative action choice rates L ∧ I and L ∧ N prevail over the detrimental action choice rates D ∧ I and D ∧ N. This effect is more pronounced for the L ∧ I choice rate than for the L ∧ N choice rate. **c** Average reversal-locked group trial-by-trial action choice rates for L ∧ I and L ∧ N actions. Both action choice rates increase over post-reversal trials. The error bars depict the SEM over reversal blocks. **d** Average group reversal-locked L ∧ I and L ∧ N choice rate difference. The difference decreased between the first trials after and the last trials before a reversal. For implementational details, please see *figure_3.m*.

the results of the model recovery analyses. For each data generating model, the corresponding panel depicts the protected exceedance probabilities (PEPs) of each data evaluation model in the model space. As shown in the leftmost panel of Figure 4a, for data generated based on the cognitive null agent C1, the PEP is maximal for the data analysis model based on agent C1, thus indicating that agent model C1 can be reliably recovered. For the data generating models based on agents C2, A1, A2, and A3, the PEPs are shown as functions of the post-decision noise parameter τ used for data generation. For data generated based on agents C2 and A1, the PEP is maximal for data analysis models based on agents C2 and A1, respectively, for all values of τ. This indicates that agent models C2 and A1 can reliably be recovered for both low and high levels of post-decision noise. For data generated based on agent A2, the PEP is maximal for the data analysis model based on agent A2 for τ values up to 0.35, while the PEP is maximal for the data analysis model based on C1, if $\tau > 0.35$. This indicates that agent model A2 can be reliably recovered for low, but not for high, levels of post-decision noise. For data generated based on agent A3 and with action valence weighting parameter set to $\lambda = 0.25$, the PEP is maximal for the data analysis model based on A3 up to a value of $\tau = 0.25$. For larger values of τ, the PEP

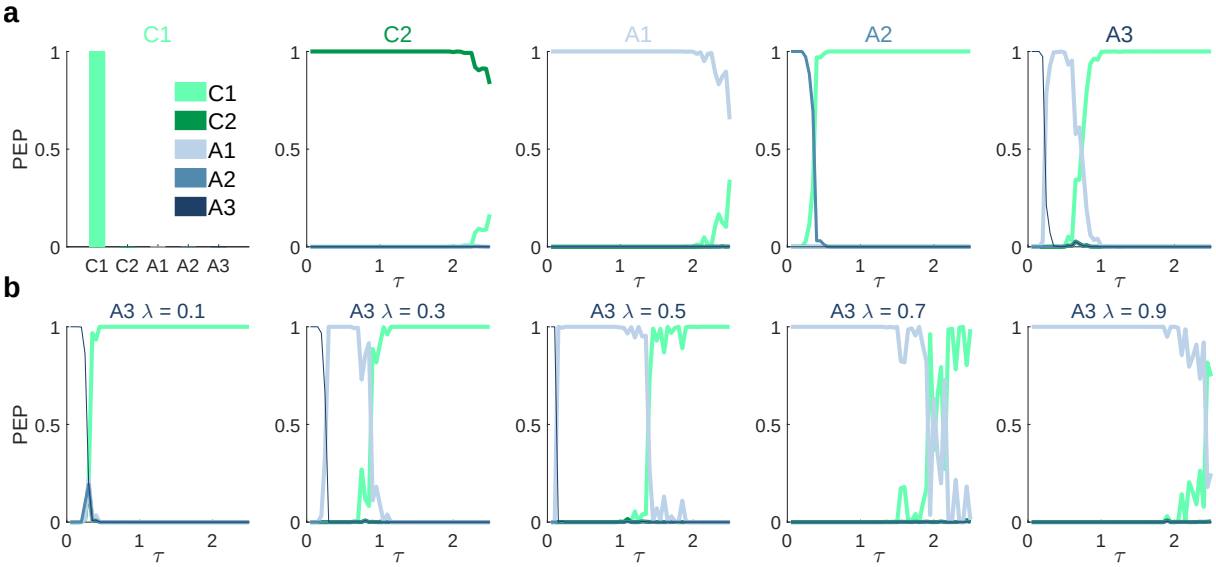

Figure 4. Model recovery results. **a** Model recovery results. Each subplot pertains to a specific data generating model (C1, C2, A1, A2, and A3) and shows the protected exceedance probabilities (PEPs) of all data analysis models in the model space evaluated in light of data sets sampled from the respective data generating model. For data generated based on the cognitive null model C1, the PEP was maximal for C1. For data generated based on agents C2, A1, A2, and A3, the PEPs depend on the value of the post-decision noise parameter τ. Agents C2 and A1 are recoverable up to high levels of post-decision noise. Agents A2 and A3 (for $\lambda = 0.25$) are recoverable only for low levels of post-decision noise. **b** Model recovery results for agent A3 for different values of the action valence weighting parameter λ. Agent A3 is recoverable up to medium values of λ and for low levels of post-decision noise. For implementational details, please see *figure_4.m*.

of the data analysis model based on agent A1 and eventually of the PEP of the data analysis model based on agent C1 exceed that of agent A3. This indicates that for $\lambda = 0.25$, the data analysis model based on agent A3 can be recovered reliably only for low levels of post-decision noise. With increasing noise, the data is better accounted for by the less complex model based on agent A1 and eventually by the cognitive null model C1.

For data generated based on agent model A3, model recovery performance depends not only on the post-decision noise parameter τ, but also on the action choice weighting parameter λ. As shown in Figure 4b, for $\lambda = 0.1$, $\lambda = 0.3$, and $\lambda = 0.5$ and for small values of τ, the PEP is maximal for the data analysis model based on agent A3. For $\lambda = 0.1$ and increasing post-decision noise, first the data analysis based on agent A2 and next the cognitive null model C1 prevail (left-most panel). For λ values of $\lambda = 0.7$ and $\lambda = 0.9$, the PEP profiles shift towards a prevailment of the model based on agent A1, in line with the increasing similarity between the action valences of agents A3 and agent A1 for increasing values of λ (cf. Figure 2c-d). More precisely, for $\lambda = 0.7$ and $\lambda = 0.9$ the PEP is maximal for model A1 up to values of $\tau = 1.9$ and $\tau = 2.4$, respectively. For even larger values of τ, the PEP of the cognitive null model C1 prevails. Together, in line with the meaning of the action valence weighting parameter of agent A3, these results indicate that agent model A3 is reliably recoverable for low to medium values of λ in the presence of low levels of post-decision noise. For very low values of λ and low levels of τ, model A3 cannot be distinguished from model A2, while for high levels of λ, it cannot be distinguished from model A1. Finally, for high levels of post-decision noise, the cognitive null model C1 provides the most parsimonious data explanation for data generated by agent model A3.

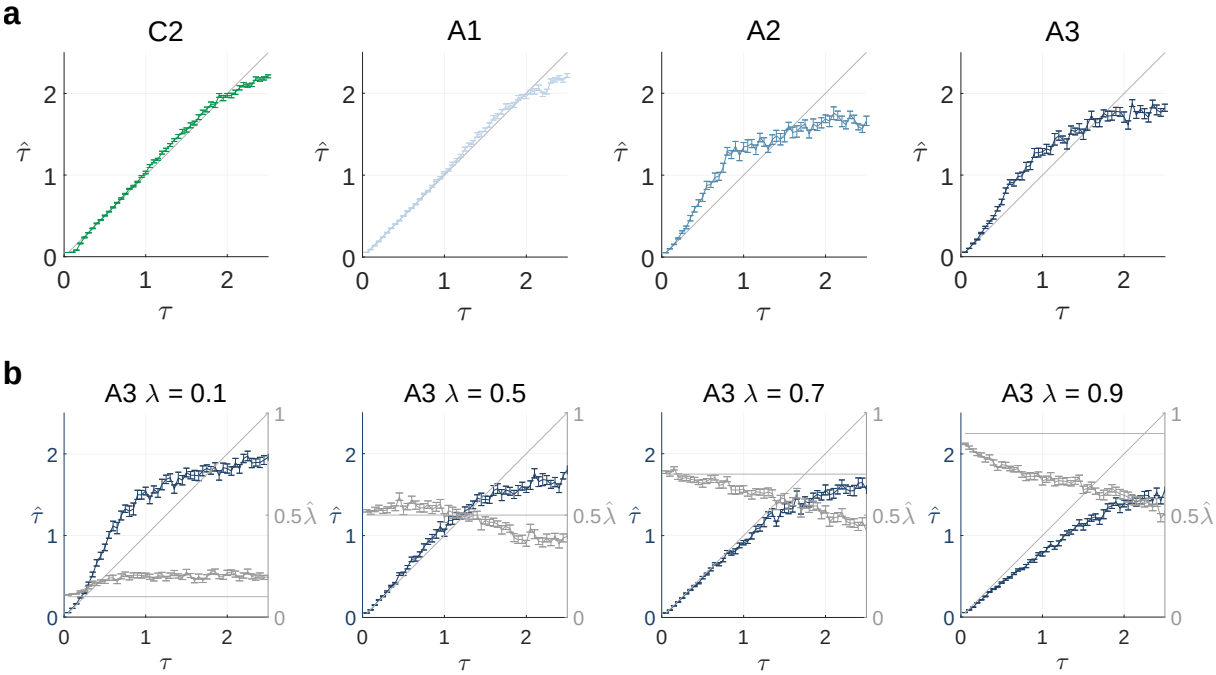

Figure 5. Parameter recovery results. a Parameter recovery results for all agents with free parameters. For every agent, the corresponding panel shows the ML parameter estimates as functions of the post-decision noise parameter τ used for data generation. The post-decision noise parameter of agents C2 and A1 is recoverable from small to medium values. The post-decision noise parameter of agents A2 and A3 with $\lambda = 0.25$ is recoverable for small values. **b** Parameter recovery results for agent A3 with different λ values. The parameter τ is recoverable for τ values up to $\tau = 1$, upon which it exhibits a downward bias. The parameter λ is recoverable for small τ values, except for a slight downward bias for $\lambda = 0.9$. For medium to large values of the post-decision noise parameter, $\hat{\lambda}$ is downward biased. For implementational details, please see *figure_ 5.m*.

Parameter recovery analyses Figure 5 summarizes the parameter recovery analyses results. Each panel of Figure 5 depicts model-specific ML parameter estimates as functions of the post-decision noise parameter τ. Specifically, Figure 5a depicts the parameter estimate $\hat{\tau}$ as a function of the respective data generating parameter values τ under the data generating and data analysis models C2, A1, A2, and A3 with $\lambda = 0.25$. In general, the parameter estimates are consistent with their respective data generating parameters for small values of τ. For larger values of τ, the true parameter values are first over- and then underestimated. As shown in Figure Figure 5a, this bias is subtle for agents C2 and A1 and only appears for relatively large τ values above $\tau = 1.5$. For agents A2 and A3 the bias is more pronounced and affects medium to large values above $\tau = 0.5$. These results are consistent with the model recovery results: For large post-decision noise, data generated by any model starts to resemble data generated by a purely random choice strategy and thus the parameter estimate for τ reaches an asymptote.

The panels of Figure 5b show the parameter estimates $\hat{\tau}$ and $\hat{\lambda}$ of agent model A3 as a function of τ for $\lambda = 0.1$, $\lambda = 0.5$, $\lambda = 0.7$, and $\lambda = 0.9$, respectively. As above, the parameter estimates $\hat{\tau}$ show a pattern of consistent estimation for values up to $\tau = 1$, upon which they exhibit a downward bias. The parameter λ is reliably recovered for small values of τ, except for a slight downward bias for $\lambda = 0.9$. However, for medium to large τ values, $\hat{\lambda}$ is more strongly downward biased. These findings are consistent with the model recovery analyses results: First, the deflation effect for $\lambda = 0.9$ and small values of τ shows that for large values of λ, data generated by agent A3 is virtually indistinguishable from data generated by agent A1 and thus $\hat{\lambda}$ reaches an asymptote. Second, the bias of $\hat{\lambda}$ for medium to large values of τ reiterates that with

increasing post-decision noise, data generated by agent A3 increasingly resembles data generated by a random choice strategy, such that λ cannot be reliably identified.

Model comparison Upon validating our modelling initiative, we evaluated and compared our agent-based model set in light of the participants' data. For 18 of the 24 participant data sets, the BIC score was maximal for the model based on agent A3. Accordingly, as shown in Figure 6a, the group cumulative BIC score was maximal for model A3 and the group-level PEP of model A3 was larger than 0.99. The ML parameter estimates $\hat{\tau}$ and $\hat{\lambda}$ for model A3 varied moderately across participants. Specifically, $\hat{\tau}$ ranged from 0.035 to 0.377 with an average of 0.124 ± 0.014 and $\hat{\lambda}$ ranged from 0.044 to 0.622 with an average of 0.274 ± 0.027. Notably, these are parameter ranges, in which A3 is well recoverable and identifiable. Taken together, these findings indicate that agent model A3 provides the most plausible explanation of the participant's choices among the set of behavioural models assessed and that the most frequently applied choice strategy among the group of participants resembled that of agent A3 most.

Post-hoc model validation Finally, we assessed the behavioural validity of the experimentally most plausible agent model A3. To this end, we first generated 24 synthetic data sets based on model A3 with parameter values set to the parameter estimates for the 24 participant data sets. We next evaluated the same summary and trial-by-trial choice rates as for the participants' data. Consistent with the empirically observed results, most synthetic action choices were lucrative and informative ($\mathrm{L} \wedge \mathrm{I}$, $84.98\% \pm 1.29$), followed by significantly fewer lucrative and non-informative synthetic actions ($\mathrm{L} \wedge \mathrm{N} 66.35\% \pm 1.96$, two-sided paired sample t-test, $t(23) = 11.59, p < 0.001$). Furthermore, as shown in Figure 6b the reward reversal-locked trial-by-trial dynamics of the synthetic actions exhibited a very similar pattern to that of the participants (cf. Figure 3c,d). Specifically, while both the average $\mathrm{L} \wedge \mathrm{I}$ as well as the average $\mathrm{L} \wedge \mathrm{N}$ action choice rates increased between two reversals (left panel), their difference decreased moderately between the first trial after and the last trial before a reward rate reversal (right panel). Taken together, these results support the behavioural validity of the A3 model. At last, we conducted model and parameter recovery analyses for model A3 based on the parameter estimates for the 24 participant data sets. As already implied by the results of the full parameter space recovery analyses reported above, these analyses confirmed that also for the range of empirically estimated parameter values, both the model and its parameters are reliably recoverable (Figure 6c and Figure 6d).

Discussion

In this work, we addressed the question of how humans make sequential decisions if all available actions bear economic consequences, but only some also deliver choice-relevant information. By collecting human choice data on an information-selective symmetric reversal bandit task, we demonstrated that in such situations, humans strive for a balance of exploratory and exploitative choice strategies. To arrive at this conclusion, we applied a set of descriptive and agent-based computational modelling analyses, including model evaluation based on formal as well as informal criteria. Given our model set, the behaviourally most plausible strategy was captured by a Bayesian agent that assessed the desirability of an action by applying a convex combination of its expected Bayesian surprise and its expected reward. A series of recovery analyses validated our modelling initiative and supports the robustness and reliability of our results. In sum, the key contributions of this work are threefold: We designed an information-selective symmetric reversal

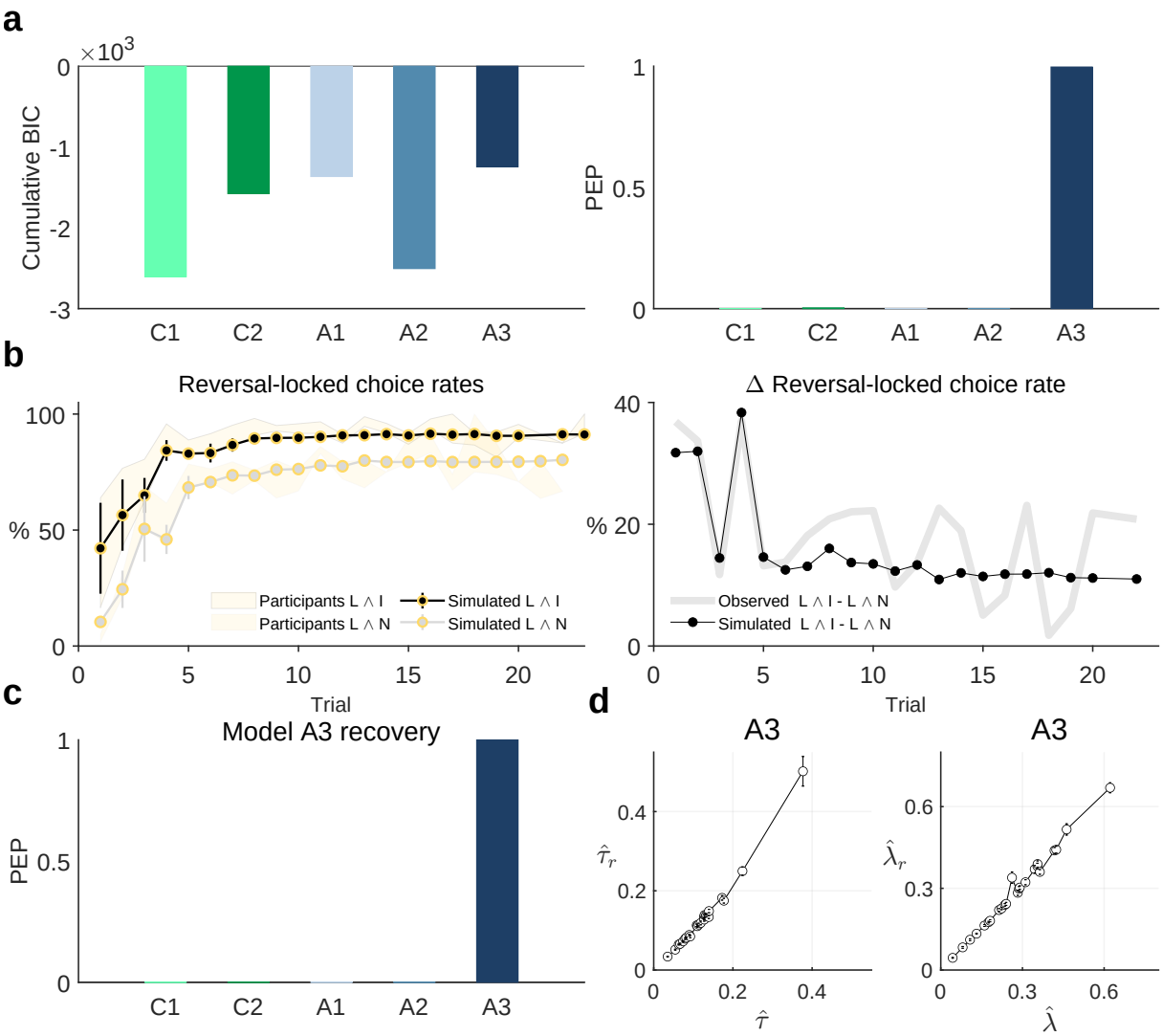

Figure 6. Computational modelling results. **a** Model comparison results. Both the group cumulative Bayesian information criterion scores (cumulative BIC, left panel) and the protected exceedance probabilities (right panel) were maximal for the agent model A3, indicating that this model explained participants' choice data the best. **b** Model A3 behavioural validation. Average reversal-locked group trial-by-trial L ∧ I action and L ∧ N action choice rates (left panel) and their difference (right panel) based on synthetic data sets generated using agent A3 and the participants' parameter estimates. The patterns closely resemble those observed in the participants' data (cf. Figures 3c, 3d), re-visualized here for convenience. **c** Model A3 recovery based on the participants' data parameter estimates. The protected exceedance probability based on re-analysis of simulated data from agent model A3 with parameters set to the participants' parameter estimates is maximal for agent model A3. **d** Parameter recovery results based on data generated with agent model A3 and the participants' data-based parameter estimates. The panels depict the simulation-averaged participant-specific recovered parameter estimates $\hat{\tau}_r$ and $\hat{\lambda}_r$ and their associated SEMs over simulations as a function of the participant-specific parameter estimates $\hat{\tau}$ and $\hat{\lambda}$. Both the participants' post-decision noise parameter estimates $\hat{\tau}$ as well as the participants' weighting parameter estimates $\hat{\lambda}$ are recoverable. For implementational details, please see *figure_ 6.m*.

bandit task, we formulated, implemented, and validated a set of agent-based behavioural models that interact with this task, and we provide empirical evidence for human uncertainty-guided exploration-exploitation choice strategies. In the following, we discuss each of these contributions in turn.

As the first of our contributions, we introduced an information-selective symmetric reversal

bandit task suitable to model a class of sequential decision-making problems, in which information about the conferred reward is not available for every action. In the following, we briefly discuss the key aspects of our task with respect to other tasks in the literature. As alluded to in the Introduction, previous research primarily employed either classical bandit paradigms or pure-exploration paradigms to study sequential decision making under uncertainty (Bubeck et al., 2009; Hertwig & Erev, 2009; Ostwald et al., 2015; Sutton & Barto, 2018; Wulff et al., 2018). We consider these paradigms in turn.

In the classical bandit paradigm, the deciding agent chooses between a set of 'arms' (Berry & Fristedt, 1985). Similar to our task, each arm confers reward according to its associated reward distribution, and, in contrast to our task, each arm confers also information about its expected reward value. A drawback of this design is that the co-occurrence of reward and information evokes a confound between an action's value estimate and the associated uncertainty: as people tend to favor the action with the higher value estimate, for that action the associated uncertainty becomes smaller - simply because for that action more reward observations are made. This makes it difficult to uncover uncertainty-guided exploration in the classical bandit paradigm (Dezza et al., 2017; Gershman, 2018; Wilson et al., 2014). The task employed in the current study circumvents this problem by adopting a symmetrical reward structure of the actions: the probability of the positive reward for the lucrative action is identical to the probability of the negative reward for the detrimental action. Likewise, the probability of the negative reward for the lucrative action is identical to the probability of the positive reward for the detrimental action. In this manner, each reward observation following the informative action confers the same amount of information about the expected reward value of both the lucrative and detrimental action. Furthermore, as in each trial information is randomly coupled with either the lucrative or the detrimental action, our task arguably evokes a more explicit exploration-exploitation dilemma than the classical bandit paradigm, in particular on trials on which participants face decisions between potentially lucrative, but non-informative and detrimental, but informative actions.

The pure-exploration paradigm models sequential decision-making problems in which an action either confers information or reward. In the classical variant of this paradigm, an action that confers reward automatically terminates the task. In an extended variant, referred to as the 'observe-or-bet task' (Blanchard & Gershman, 2018; Navarro et al., 2016; Tversky & Edwards, 1966), the deciding agent can freely switch between an 'observe action' that confers information and 'bet actions' that confer reward. Specifically, 'observe actions' return information about the expected reward of the bet actions, but do not return reward. 'Bet actions', on the other hand, return rewards according to their associated reward distributions, but no information. Similar to the bet actions, in our task one of the available actions confers only reward, but no information. However, in our task, the other available action does not only confer information, as the 'observe action' does, but it also confers reward. Therefore, while exploration and exploitation are temporally separated in the observer-or-bet task, they have to be balanced simultaneously in our task.

In summary, the task proposed in the current study complements the experimental arsenal for studying human exploration-exploitation behavior in the following sense: In contrast to the classical bandit paradigm, the exploration-exploitation dilemma on each trial is *explicit*, i.e., participants need to actively decide whether to gather reward and/or information, and do not obtain information as a by-product of reward-maximizing decisions as in the classical bandit task. Further, in contrast to the pure-exploration and observe-or-bet task, exploration and exploitation phases occur *simultaneously* and not separated in time. Finally, by introducing a constant probability for the reward rate reversal of the choice options, the task does not only evoke an explicit and simultaneous exploration-exploitation dilemma, but also one that is

ongoing. This aspect is in line with recent studies using classical bandit or observe-or-bet tasks that also adopted non-stationary reward structures (Bartolo & Averbeck, 2020; Blanchard & Gershman, 2018; Chakroun et al., 2020; Navarro et al., 2016; Hampton et al., 2006; Speekenbrink & Konstantinidis, 2015) and emulates the ever-changing reward structure of real environments.

Our second contribution is the formulation and validation of a set of agent-based behavioural models that can interact with the information-selective symmetric reversal bandit task. Specifically, our model space comprises belief state-based agents that formalize subjective uncertainty-based exploitative, explorative and hybrid explorative-exploitative strategies, as well as belief state-free agents that formalize random choice and heuristic win-stay-lose-switch strategies. The belief state-based agents implement recursive belief-state updates to infer the not directly observable structure of the task environment, i.e., the most likely currently lucrative action. While this form of optimal Bayesian learning may seem to be a strong assumption, it has been shown to approximate human learning reasonably well in comparable switching state tasks, such as the two-armed reversal bandit task, non-stationary versions of the observe-or-bet task, and the multi-armed bandit task (Hampton et al., 2006; Blanchard & Gershman, 2018; Navarro et al., 2016; Daw et al., 2006; Speekenbrink & Konstantinidis, 2015). Moreover, by also including belief state-free agents in our model space, we also accounted for simple strategies that do not require Bayesian update. Of these, the win-stay-lose-switch agent adopts a well established effective strategy to approach similar bandit problems (Robbins, 1952).

The three belief state-based agents implement their respective choice strategies by following different optimization aims. In particular, the belief state-based exploitative agent A1 seeks to maximize the belief state-weighted expected reward. The belief state-based explorative agent A2 seeks to maximize the expected Bayesian surprise. The belief state-based hybrid explorative-exploitative agent A3 seeks to maximize the convex combination of these two quantities. Belief state-weighted expected reward is a natural quantity to formally capture immediate reward gain and thus myopic exploitation (Sutton & Barto, 2018). Expected Bayesian surprise is one of many quantities that have been proposed to formally capture immediate information gain and thus myopic exploration (Schwartenbeck et al., 2019). As alluded to in the Introduction, we here opted for Bayesian surprise due to its putative representation in the human neurocognitive system (Itti & Baldi, 2009; Ostwald et al., 2012; Gijsen et al., 2020).

Importantly, we would like to note that our definition of exploration here pertains to a form of exploration that is generally referred to as 'directed exploration' (Gershman, 2018, 2019; Wilson et al., 2014). This term is used to distinguish information gain maximizing exploration from 'random exploration'. Random exploration is a form of exploration that achieves information gain by implementing a stochastic strategy, i.e. making stochastic choices based on the actions' reward value estimate. While there are more principled ways such as Thompson sampling (Thompson, 1933), random exploration is commonly accounted for by the softmax operation (Reverdy & Leonard, 2015). Notably, we here do not interpret the softmax operation as random exploration. Instead, we employ the softmax operation to embed the agents in a statistical framework that accounts for post-decision noise. This way, we clearly separate the (deterministic) choice strategies implemented by the agents and the statistical agent-based behavioural data analysis models. In future work, we aim to broaden our model space by considering agents that adopt random exploration. Crucially, this necessitates a probabilistic filtering framework that allows to partition the variability of choice data into components relating to an agent's stochastic exploration strategy and to post-decision noise (c.f. Ostwald et al., 2014).

As a third contribution, we provided evidence for human uncertainty-guided exploration-exploitation in the information-selective symmetric reversal bandit task: As uncertainty decreased, participants' choices were less influenced by the prospect of information gain and more

influenced by the prospect of reward gain. This finding is consistent with the behavior in the observer-or-bet task. In the first empirical study using the observer-or-bet task, (Tversky & Edwards, 1966) found that participants explored more, i.e., chose the observe action more frequently, if they (falsely) believed that the underlying environment was dynamic, i.e., the lucrative and detrimental bet actions reversed over time. While (Tversky & Edwards, 1966) did not relate this result to the notion that dynamic environments promote uncertainty, which, in turn, promotes exploration, in a recent study, (Navarro et al., 2016) formally tested this hypothesis. By modelling participants' choices in both static and dynamic versions of the observer-or-bet task, they demonstrated that switches between the exploratory observe action and the exploitative bet actions were mediated by uncertainty. Our result is also in line with recent findings from studies employing multi-armed bandit tasks. Work by several authors showed that when controlling for the value estimate-uncertainty confound, behavior in static two-armed bandit tasks reflects an uncertainty-dependent combination of exploration and exploitation (Dezza et al., 2017; Gershman, 2018, 2019; Wilson et al., 2014). Notably, however, consistent with the notion that the value estimate-uncertainty confound has the potential to mask directed exploration, findings from earlier studies not accounting for this confounding effect are less conclusive. For example, (Zhang & Angela, 2013) also found evidence for a belief state-based explorative-exploitative strategy in static four-armed bandit tasks. In contrast, (Daw et al., 2006) did not find evidence for such a strategy in analyzing choices in a dynamic four-armed bandit task with changing action values. While the finding from (Daw et al., 2006) is contrary to our finding, acknowledging that value estimate and uncertainty are not confounded in our task, these two findings can be reconciled.

At last, some limitations of our study along with some suggestions for future research must be acknowledged. First, as for any scientific study, inferences can only be made with respect to the model space, which *per se* is incomplete. Thus, although we provide evidence for a strategy resembling a belief state-based agent seeking to maximize the convex combination of its expected Bayesian surprise and its expected reward, it is very well possible that an agent not included in our model space can better account for the behavioral data. Based on our experimental documentation and the open availability of the data, future research may broaden the model space and could also consider, for example, probabilistic variants of the win-stay-lose-switch agents (Worthy et al., 2012) or agents with non-adaptive learning rates (Wiering & Otterlo, 2012; Rescorla & Wagner, 1972), besides the already mentioned random exploration agents in a probabilistic filtering framework (Ostwald et al., 2014). Second, on a related note, we here do not derive an optimal agent in the sense of partially observable Markov decision process theory (Bertsekas, 2000; Puterman, 2005; Bäuerle & Rieder, 2011). A model space that comprises an agent implementing an optimal choice strategy would afford the analysis of the participants' behavior from a normative perspective and is thus an interesting endeavor for future research. To facilitate such an undertaking, we have provided a detailed documentation of the task and agents developed thus far. Third, given the modest sample size of our study, the behavioral data reported here are best capitalized on in an exploratory fashion. Together with the implemented open research measures (Ritchie, 2020), we believe to have laid the foundations for reproducing our study and build upon our work.

Conclusion

In conclusion, we here introduced a new behavioural task that models a subset of real-life sequential decision-making problems that has thus far received relatively little attention in the computational modelling literature: problems, in which information about the conferred reward is not available for every action. Importantly, this task allows to investigate a pronounced form

of simultaneous exploration and exploitation processes without introducing a value estimation-uncertainty confound. By analyzing participants' choices on this task using descriptive methods and agent-based behavioural models, we provide evidence for an uncertainty-guided balance between exploration and exploitation in human sequential decision making under uncertainty.

Acknowledgements

We are grateful to Russell Tobe, Melissa Breland, Tina Bermudez, and Alexis Lieval for help with participant recruitment, Anna MacKay-Brandt for help with participant enrollment, and Caixia Hu and Raj Sangoi for help with data acquisition. We thank Hauke Heekeren for support throughout this work.

Declarations

Funding The research was supported by an Elsa-Neumann-Stipendium des Landes Berlin (L.H.).

Conflicts of interest/Competing interests Non declared.

Ethics approval The study was conducted in line with the human participant guidelines of the 1964 Declaration of Helsinki and was approved by the Institutional Review Board of the Nathan Kline Institute for Psychiatric Research (Orangeburg, NY, US).

Consent to participate and consent for publication Informed consent for participation and publication was obtained from all individual participants included in the study.

Availability of data and material The behavioural datasets generated and analysed during the current study are available from the Open Science Framework at `https://osf.io/vdmah/`.

Code availability All custom Python and Matlab code implementing the experimental paradigm, simulations, and analyses is available from the Open Science Framework at `https://osf.io/vdmah/`.

Authors' contributions L.H.: conceptualization, project administration, data curation, investigation, methodology, formal analysis, software, validation, visualization, writing – original draft, writing – review & editing; S.C.: project administration, data curation, investigation, experimental software, funding acquisition, resources, supervision, writing – review; M.M.: conceptualization, funding acquisition, resources, supervision, project administration, writing – review; S.R.: project administration, data curation, investigation, writing – review; P.S.: conceptualization, writing – review; D.O.: conceptualization, formal analysis, methodology, software, visualization, project administration, resources, validation, funding acquisition, supervision, writing – original draft, writing – review & editing.

References

Bartolo, R. & Averbeck, B. B. (2020). Prefrontal cortex predicts state switches during reversal learning. *Neuron*.

Bäuerle, N. & Rieder, U. (2011). *Markov Decision Processes with Applications to Finance*. Universitext. Berlin, Heidelberg: Springer Berlin Heidelberg.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Beck depression inventory-II. *San Antonio*, 78(2), 490–498.

Berry, D. A. & Fristedt, B. (1985). Bandit problems: Sequential allocation of experiments (Monographs on statistics and applied probability). *London: Chapman and Hall*, 5, 71–87.

Bertsekas, D. P. (2000). *Dynamic Programming and Optimal Control*. Athena Scientific, second edition.

Blanchard, T. C. & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, 18(1), 117–126.

Brand, H. & Woods, P. J. (1957). Effects of a random versus pattern instructional set in a contingent partial reinforcement situation. *Psychological Reports*, 3, 473–479.

Brand, H., Woods, P. J., & Sakoda, J. M. (1956). Anticipation of reward as a function of partial reinforcement. *Journal of Experimental Psychology*, 52(1), 18–22.

Bubeck, S., Munos, R., & Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory* (pp. 23–37).: Springer.

Byrd, R. H., Gilbert, J. C., & Nocedal, J. (2000). A trust region method based on interior point techniques for nonlinear programming. *Mathematical programming*, 89(1), 149–185.

Byrd, R. H., Hribar, M. E., & Nocedal, J. (1999). An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4), 877–900.

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, 9, e51260.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481), 933–942.

Conners, C. K., Erhardt, D., & Sparrow, E. (1999). CAARS. Adult ADHD rating scales. *Technical manual.*

Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A., & Averbeck, B. B. (2016). Amygdala and ventral striatum make distinct contributions to reinforcement learning. *Neuron*, 92(2), 505–517.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095), 876–879.

Dayan, P. & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4), 429–453.

Dezza, I. C., Angela, J. Y., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific reports*, 7(1), 1–13.

Even-Dar, E., Mannor, S., & Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. (pp.$\tilde{2}$7).

Gabillon, V., Ghavamzadeh, M., & Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. (pp.$\tilde{9}$).

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

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, 6(3), 277–286.

Gijsen, S., Grundei, M., Lange, R. T., Ostwald, D., & Blankenburg, F. (2020). Neural surprise in somatosensory Bayesian learning. *bioRxiv.*

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral cortex*, 19(2), 483–495.

Hampton, A. N., Bossaerts, P., & O'doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26(32), 8360–8367.

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *Neuroimage*, 84, 159–168.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese*, 187(1), 269–292.

Hertwig, R. & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, 13(12), 517–523.

Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision research*, 49(10), 1295–1306.

Kullback, S. & Leibler, R. A. (1951-03, 1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, 12(2), 164–174.

Lynam, D. R., Smith, G. T., Whiteside, S. P., & Cyders, M. A. (2006). The UPPS-P: Assessing five personality pathways to impulsive behavior. *West Lafayette, IN: Purdue University.*

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive psychology*, 85, 43–77.

Nooner, K. B., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., Panek, L., Brown, S., Zavitz, S., Li, Q., et al. (2012). The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, 6, 152.

Ostwald, D., Kirilina, E., Starke, L., & Blankenburg, F. (2014). A tutorial on variational Bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, 60, 1–19.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*, 62(1), 177–188.

Ostwald, D., Starke, L., & Hertwig, R. (2015). A normative inference approach for optimal sample sizes in decisions from experience. *Frontiers in psychology*, 6, 1342.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, 162(1-2), 8–13.

Puterman, M. (2005). *Markov Decision Processes.* Wiley-Interscience.

Rescorla, R. A. & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In *Classical Conditioning II* (pp.1̄8). New York.

Reverdy, P. & Leonard, N. E. (2015). Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering*, 13(1), 54–67.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, 84, 971–985.

Ritchie, S. (2020). *Science Fictions - Exposing Fraud, Bias, Negligence and Hype in Science.* Penguin.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5), 527–535.

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, 8, e41703.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461–464.

Speekenbrink, M. & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, 7(2), 351–367.

Spielberger, C., Gorsuch, R., Lushene, R., Vagg, P., & Jacobs, G. (1983). Manual for the state-trait anxiety inventory (palo alto, CA, consulting psychologists press). *Inc.*

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, 46(4), 1004–1017.

Sun, Y., Gomez, F., & Schmidhuber, J. (2011). Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence* (pp. 41–51).: Springer.

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning: An Introduction.* MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4), 285–294.

Tversky, A. & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, 71(5), 680.

Waltz, R. A., Morales, J. L., Nocedal, J., & Orban, D. (2006). An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical programming*, 107(3), 391–408.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence.* The Psychological Corporation: Harcourt Brace & Company. New York, NY.

Wiering, M.-H. & Otterlo, Y.-S. (2012). *Reinforcement Learning.*

Wilson, R. C. & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, 8, e49547.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6), 2074.

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(6), 1640–1658.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12), 915–924.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin*, 144(2), 140.

Zhang, S. & Angela, J. Y. (2013). Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in Neural Information Processing Systems* (pp. 2607–2615).

Supplementary material

S.1. Sample characteristics

To characterize the group of participants, we measured symptoms of attention deficit hyperactivity disorder (ADHD), anxiety, depression and impulsivity. To this end, we used the questionnaires Conners Adult ADHD Rating Scale – Self Report, Short Version (CAARS-S:S; Conners et al. (1999)), State and Trait Anxiety Inventory (STAI; Spielberger et al. (1983)), Beck Depression Inventory II (BDI-II; Beck et al. (1996)) and UPPS-P Impulsive Behavior Scale (Lynam et al., 2006), respectively. As shown in Table S.1, the sample varied only moderately with respect to these symptoms. For example, on CAARS-S:S, our main questionnaire of interest, participants scored within ±2 standard deviations of the mean of their age- and gender-matched norm groups of the general population. We therefore concluded that the sample represents the healthy population and did not relate individual variability in terms of ADHD or other clinical symptoms to behavioural strategies. We also report the IQ score, which was obtained by administering the Wechsler Abbreviated Scale of Intelligence (WASI-II; Wechsler (1999)) at the time of the Nathan Kline Institute Rockland Sample study (Nooner et al., 2012).

Measurement	**Range**	**Median**	**Mean ± SD**
Age (years)	18 - 35	23.5	24.5 ± 5.53
WASI-II (total score)	84 - 122	101.5	102.38 ± 9.14
CAARS-S:S (total T-score)	32 - 65	48	47.63 ± 9.16
BDI-II (total)	0 - 20	4	6.67 ± 6.3
STAI STATE (total T-score)	34 - 63	43	45.6 ± 8.52
STAI TRAIT (total T-score)	34 - 71	49.5	49.92 ± 10.29
UPPS-P (total)	73 - 190	118.5	124.04 ± 25.66

Table S.1. Sample characteristics.

S.2. Participant instructions

Participants were provided with the following task instructions:

Welcome to the main part of today's experiment! In the following we will introduce to you the decision making task that you will complete in the scanner. Please read the instructions carefully. If you have any questions, feel free to ask at any time. Once you read the instructions, you will complete a test run with the task to make sure you feel comfortable with it before going in the scanner. On every trial we will present to you two objects, an orange square and a blue triangle on either side of a black and grey screen and ask you to choose between them. One of these objects is profitable, meaning that it is going to give you a win most of the time, while the other object in not profitable meaning that it is going to give you a loss most of the time. Once you choose one of the objects, the outcome (win: +1 or loss: −1) will be registered to your account. You will have 2.5 seconds to indicate your choice. If you do not respond within this time window, the message 'Too slow' will appear on the screen and you automatically lose 1 point. Here you see an example for a trial (Supplementary Figure S.1).

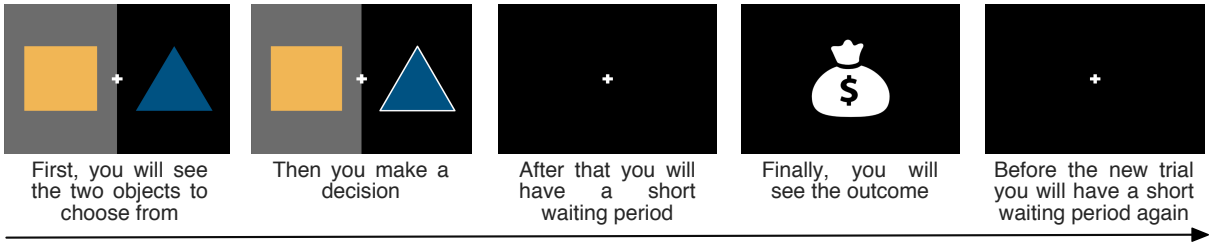

Figure S.1. Participant instructions 1. The figure shows the sequence of events within a trial as presented to the participants in the instructions.

You will start the experiment with a balance of 0 points and any wins or losses will be registered to your account. After the experiment, in addition to your standard payment for participation, you will receive up to $30 depending on your final account. Note that your balance cannot get below 0 and if you do not earn additional money on the task, you will not be penalized and you will still receive the standard payment for your participation. We would however encourage you to try to earn as much as possible on the task. After each run we will show you your balance. A run consists of 80 trials, which takes about 20 minutes to complete. You will have two runs in the scanner.

As mentioned above, one of the objects is profitable and it will bring you a win most of the times and every now and then it will bring you loss. At the same time, the other object is not profitable and it will bring you a loss most of the time and a win every now and then. You won't explicitly know which object is the profitable one and which is the non-profitable and you will need to conclude it from the outcomes. But be aware! These roles can switch, which means that the previously profitable object becomes non-profitable and the previously non-profitable object becomes profitable. Such a switch will happen only 1-4 times in the entire run and you will have enough trials without a switch to conclude which object is the profitable one.

Keep in mind that even the currently profitable object can from time to time deliver a loss and a couple of negative outcomes does not necessarily mean that a switch occurred. Similarly, even the non-profitable object can from time to time deliver a win and a couple of positive outcomes does not necessarily mean that a switch occurred. You can however assume that a switch has happened if you feel the previously rewarding object started to give you more losses than wins and the previously non-rewarding object started to bring you more wins than losses.

Before you do the test run, there is one more important aspect to the task: On each trial, one of the objects will be presented to you in front of a black background while the other object will be in front of a grey background. If you choose the object on the black-side, you will see the outcome of your choice. However, if you choose the grey-side object, the outcome will remain hidden from you but it will be registered to your account (Supplementary Figure S.2).

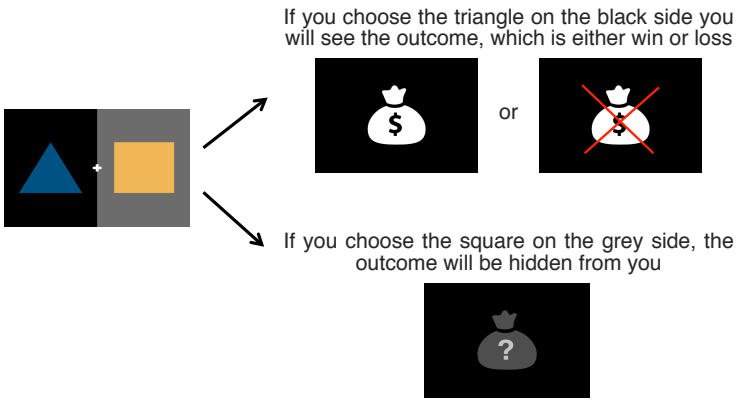

Figure S.2. Participant instructions 2. The figure depicts the lucrativeness and informativeness associated with the actions as presented to the participants in the instructions.

You will now complete a test run, which will be just like the ones you will complete in the scanner. We will discuss all your questions to make sure you feel comfortable with the task before going in the scanner.

Note that from the perspective of the participants the specification of the available actions was doubly over-specified: on the one hand, the square and the triangle were also colored orange and blue ,and on the other hand, one side of the screen was also indicated by a black background, while the other side was indicated in by a gray background. For efficiency, in the main text, we only retain the the notions of squares and triangles for available rewarded actions, black and grey backgrounds for informative and non-informative actions, and we use colors to indicate the currently lucrative (yellow) and detrimental (blue) actions.

S.3. Experimental state sequence

Table S.2 displays the definition of the state evolution function f. The upper table displays the state sequence of the first experimental run, the lower table displays the state sequence of the second run. t indexes the trial number and s_t is the trial state.

t	1	2	3	4	5	6	7	8	9	10
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$
t	11	12	13	14	15	16	17	18	19	20
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(2,1)$
t	21	22	23	24	25	26	27	28	29	30
s_t	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$
t	31	32	33	34	35	36	37	38	39	40
s_t	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$
t	41	42	43	44	45	46	47	48	49	50
s_t	$(2,2)$	$(2,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,2)$
t	51	52	53	54	55	56	57	58	59	60
s_t	$(1,1)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$
t	61	62	63	64	65	66	67	68	69	70
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,1)$
t	71	72	73	74	75	76	77	78	79	80
s_t	$(2,2)$	$(2,1)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$

t	1	2	3	4	5	6	7	8	9	10
s_t	$(2,1)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,1)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,2)$
t	11	12	13	14	15	16	17	18	19	20
s_t	$(2,1)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$
t	21	22	23	24	25	26	27	28	29	30
s_t	$(2,1)$	$(2,1)$	$(2,2)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,1)$
t	31	32	33	34	35	36	37	38	39	40
s_t	$(1,2)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$
t	41	42	43	44	45	46	47	48	49	50
s_t	$(1,2)$	$(1,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$
t	51	52	53	54	55	56	57	58	59	60
s_t	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(1,2)$
t	61	62	63	64	65	66	67	68	69	70
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$
t	71	72	73	74	75	76	77	78	79	80
s_t	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,2)$

Table S.2. Experimental state sequence.

S.4. Belief state, posterior predictive distribution, and KL-divergence

Belief state

Based on the probability distributions $p\left(s_1^1\right)$, $p\left(s_{t+1}^1|s_t^1\right)$ and $p^{a_t}\left(o_t|s_t^1\right)$, the agent's belief state on trials $t = 2, ..., T$ can be recursively evaluated according to eq. (15). To show the validity of this equation, we first express the belief state as

$$p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) = \frac{p^{a_{1:t-1}}\left(s_t^1, o_{1:t-1}\right)}{p^{a_{1:t-1}}\left(o_{1:t-1}\right)}. \tag{S.1}$$

The numerator of eq. (S.1) can then be rewritten as

$$\begin{aligned}
p^{a_{1:t-1}}\left(s_t^1, o_{1:t-1}\right) &= \sum_{s_{t-1}^1} p^{a_{1:t-1}}\left(s_t^1, s_{t-1}^1, o_{1:t-1}\right) \\
&= \sum_{s_{t-1}^1} p^{a_{1:t-1}}\left(s_t^1|s_{t-1}^1, o_{1:t-1}\right) p^{a_{1:t-1}}\left(s_{t-1}^1, o_{1:t-2}, o_{t-1}\right) \\
&= \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-2:t-1}}\left(o_{t-1}|s_{t-1}^1, o_{1:t-2}\right) p^{a_{1:t-2}}\left(s_{t-1}^1, o_{1:t-2}\right) \\
&= \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right) p^{a_{1:t-2}}\left(o_{1:t-2}\right) \\
&= p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right),
\end{aligned} \tag{S.2}$$

where we used the conditional independence of s_t^1 of $o_{1:t-1}$ given s_{t-1}^1 in the third equality and the conditional independence of o_{t-1} of all other random variables given s_{t-1}^1 in the fourth equality. Similarly, we can rewrite the denominator of eq. (S.1) as

$$\begin{aligned}
p^{a_{1:t-1}}\left(o_{1:t-1}\right) &= \sum_{s_t^1} \sum_{s_{t-1}^1} p^{a_{1:t-1}}\left(s_t^1, s_{t-1}^1, o_{1:t-1}\right) \\
&= p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_t^1} \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right),
\end{aligned} \tag{S.3}$$

where in the last equality we used the numerator's form derived in eq. (S.2). Finally, by substitution of (S.2) and (S.3) in (S.1), we obtain the belief state update formula of eq. (15) as follows:

$$\begin{aligned}
p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) &= \frac{p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}{p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_t^1} \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)} \\
&= \frac{\sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}{\sum_{s_t^1} \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}.
\end{aligned} \tag{S.4}$$

Posterior predictive distribution

Given the agent's belief state $p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right)$ and the action-dependent state-conditional observation distribution $p^{a_t}\left(o_t|s_t^1\right)$, the posterior predictive distribution can be evaluated according to eq. (23). A proof of this equation is as follows:

$$\begin{aligned}
p^{a_{1:t}}\left(o_t|o_{1:t-1}\right) &= \frac{p^{a_{1:t}}\left(o_t, o_{1:t-1}\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t}}\left(o_t, o_{1:t-1}, s_t^1\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t}}\left(s_t^1|o_t, o_{1:t-1}\right) p^{a_{1:t}}\left(o_t, o_{1:t-1}\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) \sum_{s_t^1} p^{a_{1:t}}\left(o_t, o_{1:t-1}, s_t^1\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) p^{a_{1:t}}\left(o_t|o_{1:t-1}, s_t^1\right) p^{a_{1:t}}\left(o_{1:t-1}, s_t^1\right)}{\sum_{s_t^1} p^{a_{1:t}}\left(o_{1:t-1}, s_t^1\right)} \\
&= \sum_{s_t^1} p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) p^{a_t}\left(o_t|s_t^1\right) \\
&= b_t p^{a_t}\left(o_t|s_t^1 = 1\right) + \left(1 - b_t\right) p^{a_t}\left(o_t|s_t^1 = 2\right), \qquad \text{(S.5)}
\end{aligned}$$

where in the last equality we substituted the belief state with its scalar representation.

KL-divergence

Recall that the KL-divergence for two distributions p and q of a discrete random variable x is defined as (Kullback & Leibler, 1951)

$$\mathrm{KL}\left(p(x)||q(x)\right) = \sum_{x \in \mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) \qquad \text{(S.6)}$$

With

$$b := p^{a_{1:t-1}}(s_t^1 = 1|o_{1:t-1}) \qquad \text{(S.7)}$$

and thus

$$p^{a_{1:t-1}}(s_t^1 = 0|o_{1:t-1}) = 1 - b, \qquad \text{(S.8)}$$

as well as

$$b^{a,o} := p^{a_{1:t-1}, a_t = a}(s_{t+1}^1 = 1|o_{1:t-1}, o_t = o) \qquad \text{(S.9)}$$

and thus

$$p^{a_{1:t-1}, a_t = a}(s_{t+1}^1 = 0|o_{1:t-1}, o_t = o) = 1 - b^{a,o} \qquad \text{(S.10)}$$

we have

$$
\begin{aligned}
&\mathrm{KL}\left(p^{a_{1:t-1},a_t=a}(s^1_{t+1}|o_{1:t-1},o_t=o)||p^{a_{1:t-1}}(s^1_t|o_{1:t-1})\right)\\
&=\sum_{s^1\in S^1} p^{a_{1:t-1},a_t=a}(s^1_{t+1}|o_{1:t-1},o_t=o)\ln\left(\frac{p^{a_{1:t-1},a_t=a}(s^1_{t+1}|o_{1:t-1},o_t=o)}{p^{a_{1:t-1}}(s^1_t|o_{1:t-1})}\right)\\
&=p^{a_{1:t-1},a_t=a}(s^1_{t+1}=0|o_{1:t-1},o_t=o)\ln\left(\frac{p^{a_{1:t-1},a_t=a}(s^1_{t+1}=0|o_{1:t-1},o_t=o)}{p^{a_{1:t-1}}(s^1_t=0|o_{1:t-1})}\right)\\
&\quad+p^{a_{1:t-1},a_t=a}(s^1_{t+1}=1|o_{1:t-1},o_t=o)\ln\left(\frac{p^{a_{1:t-1},a_t=a}(s^1_{t+1}=1|o_{1:t-1},o_t=o)}{p^{a_{1:t-1}}(s^1_t=1|o_{1:t-1})}\right)\\
&=(1-b^{a,o})\ln\left(\frac{1-b^{a,o}}{1-b}\right)+b^{a,o}\ln\left(\frac{b^{a,o}}{b}\right)
\end{aligned}
\tag{S.11}
$$

S.5. Belief state and posterior predictive distribution implementation

For a concise implementation of the belief state update and the posterior predictive distribution formulas, we represent the agent's probability distributions by stochastic vectors and stochastic matrices. Specifically, in the implementation of the agent model components as defined in *irb_modcomp.m*,

- $\mu_1 \in \mathbb{R}_{\geq 0}^{|S^1|}$ represents the initial belief state $p\left(s_1^1\right)$. The ith entry of μ_1 corresponds to the agent's subjective uncertainty that the non-observable state component takes on value $s_1^1 = i$ at trial $t = 1$. Formally,

$$\mu_1 := \begin{pmatrix} p\left(s_1^1 = 1\right) \\ p\left(s_1^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}. \tag{S.12}$$

- $\mu_t \in \mathbb{R}_{\geq 0}^{|S^1|}$ represents the belief state $p^{a_{1:t-1}}\left(s_t^1 | o_{1:t-1}\right)$ at trial t. The ith entry of μ_t corresponds to the agent's subjective uncertainty that the non-observable state component takes on value $s_t^1 = i$ at trial t given the history of observations $o_{1:t-1}$ and actions $a_{1:t-1}$. Formally,

$$\mu_t := \begin{pmatrix} p^{a_{1:t-1}}\left(s_t^1 = 1 | o_{1:t-1}\right) \\ p^{a_{1:t-1}}\left(s_t^1 = 2 | o_{1:t-1}\right) \end{pmatrix}. \tag{S.13}$$

- $\Phi \in \mathbb{R}_{\geq 0}^{|S^1| \times |S^1|}$ represents the state-state transition distribution $p\left(s_{t+1}^1 | s_t^1\right)$. The jth entry of the ith row of Φ corresponds to the agent's subjective uncertainty that the non-observable state component takes on the value $s_{t+1}^1 = j$ in trial $t+1$ given that $s_t^1 = i$ in trial t. Formally,

$$\Phi := \begin{pmatrix} p\left(s_{t+1}^1 = 1 | s_t^1 = 1\right) & p\left(s_{t+1}^1 = 2 | s_t^1 = 1\right) \\ p\left(s_{t+1}^1 = 1 | s_t^1 = 2\right) & p\left(s_{t+1}^1 = 2 | s_t^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0.9625 & 0.0375 \\ 0.0375 & 0.9625 \end{pmatrix}. \tag{S.14}$$

- $\Omega^{a_t} \in \mathbb{R}_{\geq 0}^{|S^1| \times |O|}$ represents the action-dependent state-conditional observation distribution $p^{a_t}\left(o_t | s_t^1\right)$ for action $a \in A$. The kth entry of the ith row of $\Omega^{a_t = a}$ corresponds to the agent's subjective uncertainty that the observation takes on the value $o_t = k$ given that the non-observable state component takes on the value $s_t^1 = i$ and the action value is $a_t = a$. Formally, for the informative actions

$$\Omega^{a_t=1} := \begin{pmatrix} p^1\left(o_t = 1 | s_t^1 = 1\right) & p^1\left(o_t = 2 | s_t^1 = 1\right) & p^1\left(o_t = 3 | s_t^1 = 1\right) \\ p^1\left(o_t = 1 | s_t^1 = 2\right) & p^1\left(o_t = 2 | s_t^1 = 2\right) & p^1\left(o_t = 3 | s_t^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0.15 & 0.85 & 0 \\ 0.85 & 0.15 & 0 \end{pmatrix}, \tag{S.15}$$

and

$$\Omega^{a_t=3} := \begin{pmatrix} p^3\left(o_t = 1 | s_t^1 = 1\right) & p^3\left(o_t = 2 | s_t^1 = 1\right) & p^3\left(o_t = 3 | s_t^1 = 1\right) \\ p^3\left(o_t = 1 | s_t^1 = 2\right) & p^3\left(o_t = 2 | s_t^1 = 2\right) & p^3\left(o_t = 3 | s_t^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0.85 & 0.15 & 0 \\ 0.15 & 0.85 & 0 \end{pmatrix}, \tag{S.16}$$

and for the non-informative actions

$$\Omega^{a_t \in \{2,4\}} := \begin{pmatrix} p^{a_t}\left(o_t = 1 | s_t^1 = 1\right) & p^{a_t}\left(o_t = 2 | s_t^1 = 1\right) & p^{a_t}\left(o_t = 3 | s_t^1 = 1\right) \\ p^{a_t}\left(o_t = 1 | s_t^1 = 2\right) & p^{a_t}\left(o_t = 2 | s_t^1 = 2\right) & p^{a_t}\left(o_t = 3 | s_t^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{S.17}$$

- $\Psi^{a_t} \in \mathbb{R}_{\geq 0}^{|S^1| \times |R|}$ represents the action-dependent state-conditional reward distribution $p^{a_t}\left(r_t | s_t^1\right)$ for action $a \in A$. The lth entry of the ith row of $\Psi^{a_t = a}$ corresponds to the agent's subjective uncertainty that the reward takes on the value $r_t = l - m$ given that the non-observable state component takes on the value $s_t^1 = i$ and action the value $a_t = a$. Note that m is introduced

to convert the linear indices to reward values and takes on the value 2 if $l = 1$ and the value 1 if $l = 2$. Formally,

$$\Psi^{a_t \in \{1,2\}} := \begin{pmatrix} p^{a_t}\left(r_t = -1 | s_t^1 = 1\right) & p^{a_t}\left(r_t = +1 | s_t^1 = 1\right) \\ p^{a_t}\left(r_t = -1 | s_t^1 = 2\right) & p^{a_t}\left(r_t = +1 | s_t^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0.15 & 0.85 \\ 0.85 & 0.15 \end{pmatrix} \tag{S.18}$$

represents the action-dependent state-conditional reward distribution for the actions of choosing the square, and

$$\Psi^{a_t \in \{3,4\}} := \begin{pmatrix} p^{a_t}\left(r_t = -1 | s_t^1 = 1\right) & p^{a_t}\left(r_t = +1 | s_t^1 = 1\right) \\ p^{a_t}\left(r_t = -1 | s_t^1 = 2\right) & p^{a_t}\left(r_t = +1 | s_t^1 = 2\right) \end{pmatrix} := \begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix} \tag{S.19}$$

represents the action-dependent state-conditional reward distribution for the actions of choosing the triangle. Accordingly, $\sum_{l=1}^{|R|} \Psi_{i(l-m)}^{a_t} = 1$.

Based on the definitions above and using the standard matrix product $\cdot$ as well as the element-wise (Hadamard) matrix product $\circ$, the agent's belief state at trial t (eq. (15))can be written as

$$\mu_t := \tilde{\mu}_t \cdot \left(\sum_{i=1}^{|S^1|} \tilde{\mu}_{t_i} \right)^{-1}, \tag{S.20}$$

where

$$\tilde{\mu}_t := \Phi \cdot (\Omega_k^a \circ \mu_{t-1}) \tag{S.21}$$

is the unnormalized belief state following action $a_{t-1} = a$ and observation $o_{t-1} = k$ and $\left(\sum_{i=1}^{|S^1|} \tilde{\mu}_{t_i}\right)^{-1}$ is the normalization constant. Here, Ω_k^a denotes the kth column of Ω^a and μ_{t-1} denotes the prior belief state on trial $t-1$, which corresponds to eq. (S.13), if $t-1 > 1$, and to eq. (S.12), if $t-1 = 1$. Similarly, the posterior predictive distribution of agent model A2 (cf. eq. (23)) for action $a_t = a$ can be written as

$$\omega_t := (\Omega^a)^T \cdot \mu_t. \tag{S.22}$$

Prof. Dr. Dirk Ostwald
Abteilung Methodenlehre I
Institut für Psychologie
Otto-von-Guericke-Universität Magdeburg

Prof. Dr. Scott Brown

Editor in Chief

Computational Brain & Behavior

Revised manuscript COBB-D-20-00031 "Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task"

Dear Prof. Dr. Scott Brown,

as encouraged by your letter of 18 November 2020, please find attached our revised manuscript (COBB-D-20-00031). We were delighted by the generally positive assessment of our work by you and two anonymous reviewers and are thankful for the detailed review and many helpful suggestions for improving the clarity and argumentative strength of the manuscript.

In light of your and the reviewers' comments, we have made the following changes to the manuscript: first, we have expanded on the real-world applicability of our laboratory task and drawn out the relationships to similar paradigms in the literature. Second, we have extensively revised the figure set of our manuscript with the aim of enhancing the accessibility of our methods and results. Finally, we have made many smaller corrections to the manuscript that improve the overall quality of our study.

Additionally, in our response letter to the reviewers, we provide a detailed account of the value and meaning of our task and agent model formulations, we explain why the suggested ANOVA analyses are not conceptually appropriate, and we go into depth to demonstrate that we in fact adhere to standard practices in reporting descriptive and inferential statistics.

Unfortunately, due to limited resources it is not possible for us to acquire additional experimental data at present. However, in the revised version of the manuscript, we emphasize the small-scale exploratory nature of our study and by implementing a comprehensive set of open research practices believe that we have laid strong foundations for others to reproduce our work. Finally, we firmly believe that the mathematically inclined and computationally well-versed target audience of *Computational Brain & Behavior* will find the level of detail and rigor exhibited by our manuscript useful for engaging with the implementation of our model-based data analyses and feel empowered to directly build upon our work, if desired.

We thank you and the reviewers for your consideration and helpful comments, and hope that you will find our revised manuscript to be suitable for publication in *Computational Brain & Behavior*.

Sincerely,

Prof. Dr. Dirk Ostwald

Reviewer #1

Thank you for the opportunity to review this paper. Clearly a lot of work has gone into this, and the rigor put into the model/parameter recovery and the math is to be commended.

We thank the reviewer for engaging with our work and their appreciation of our scientific rigour.

However, while the math in the paper is quite rigourous (but see the comment on Eq. 29), relatively simple ideas are explained with excessive complexity. For instance, the random choice agent (C1) requires going through Eqs. 4 - 12 in addition to Eq. 27. This is not merely a minor issue, but makes large swathes of the paper incredibly difficult to comprehend, despite the relative simplicity of the models. I understand that this journal places a strong weight on rigour, yet this level of complexity hurts the readibility of this paper just as much as an overfit regression model might hurt the interpretability of the results. I would strongly recommend that the authors consider an alternative formulation of the models in the paper and a substantial reduction in notational overload.

Here is one potential solution, although feel free to consider alternatives: 1. Remove the task model from the main text and move to the SI. The description of the experiment was sufficient.

We thank the reviewer for this suggestion, however, from our point of view, a specification of the task model in the main text - beyond the description of the experiment - is essential for the following reasons. First, while the reviewer may feel that our verbose description of the experiment is sufficient to equip the reader with a good understanding of the experimental methods, a verbose description does not suffice to implement agent models that act in an environment with the same statistical regularities as constructed in the experiment. Clearly, agent models require a formal specification and core structural aspects of the agent models correspond to aspects of the task model (such as the state, action, and observation spaces as well as the state-action-conditional reward distribution). This reflects the assumption, that the decision makers (agent models/participants) encode knowledge of certain aspects of the experiment (such as the reward structure of the shapes). Without a formalized task model, the specification of this agent model knowledge is not possible. In other words, we argue that for a full specification of the agent models, a specification of the task model is essential in the main text. Second, the specification of a task model is also a necessary prerequisite for the implementation of agent model simulations. To underline this argument, we invite the reviewer to engage with the task model implementation in *irb_taskobj.m*, which incorporates a clear documentation in terms of the task model.

2. Rather than describing a general model that covers all agent types, find a minimally sufficient description for each agent that allows you to highlight the key differences.

The reviewer raises the important issue of providing a minimally sufficient description of each agent. The minimally sufficient description of each agent is its action valence function, as the agent's action valence function specifies how a value-maximizing agent values its available actions on trial t based on a history of trials $1 : t - 1$. Differences in action valence functions thus correspond to key differences in agents. In our original submission, these are provided in eq. (12) for Agent C1, eqs. (12) and (14) for Agent C2, eq. (19) for Agent A1, eq. (22) for Agent A2, and eq. (25) for Agent A3.

For instance, none of the control models need a description of beliefs states, since they are omitted. Simply define a random choice policy and the case-based policy of the WSLS agent.

We thank the reviewer for this comment and agree that the control models do not encode belief states. This is why they are defined before the introduction of the belief state mechanics of the Bayesian agents and were simply defined by a random choice action valence function (eq. (12)) and the WSLS action valence function (eqs. (13) and (14)). As special cases of the general agent model M_{Agent} and as opposed to the Bayesian agent set, they hence capitalize only on the definitions of T, S, A, R and O.

At no point does the 9 component tuple in Eq. 4 provide any benefit.

The tuple M_{Agent} in eq. (4) defines the structure and core aspects of the functionality of the agent models

used to explain the participant's choice data. Without the definition of M_{Agent}, the structure and function of the agent models cannot be communicated precisely. We expand on the theoretical role and importance of the tuple M_{Agent} in relation to the reviewer's second comment on eq. (4) below. We also note that the tuple, and, more importantly, the ensuing definitions of its components on pp. 9 - 10, conform to a verbose documentation and explanation of the respective programming entities that we developed in the computational implementation of our modeling initiative. To underline this argument, we invite the reviewer to inspect the Matlab functions *irb_modcomp.m* and the agent class definition files *irb_agentobj_*.m*, which incorporate a clear documentation in terms of the general agent model.

3. Simplify notation. Is the s^1 superscript really necessary when it never takes any other values?

We would like to take this remark as an opportunity to elucidate our notation of the state variable. As described on p. 7, lines 230 - 237, the s^1 superscript 1 indexes the first component of the two-dimensional state outcome values $s = (s^1, s^2) \in \mathbb{N}_2 \times \mathbb{N}_2$. Indexing vector components by numbers is standard mathematical notation. A two-dimensional state vector is necessary here as there are unobservable and observable state components (the lucrative shape and the available actions, respectively). Replacing numerical indexing with Roman letter variations (e.g. $s := (x, y)$) would required more symbols and make it less apparent that x and y refer to state vector components. Moreover, the random variables $s^1_t, t = 1, ..., T$ take on values $s^1 \in \mathbb{N}_2$. As detailed on p. 7, lines 230 - 237, s^1_t takes on the value 1 if the square is lucrative and s^1_t takes on the value 2, if the triangle is lucrative. As described on p. 4, lines 134 - 140, there were three shape lucrativeness reversals per task run. The detailed sequence of s^1_t values is documented in Supplementary Material S.3. For example, $s^1_{19} = 1$ and $s^1_{20} = 2$. Thus, s^1_t does in fact take on different values. However, s^1 always refers to the first (unobservable) state component, hence the superscript 1.

What's the point of introducing the 9 component tuple describing the general agents, when very few of the components are actually necessary for explaining your belief-based agents in Eqs. 19 - 24?

To answer the reviewer's questions, we first recapitulate that the general agent structure (referred to as the 9 component tuple by the reviewer) is given by

$$M_{\text{Agent}} := (T, S, A, R, O, p(s^1_1), p(s^1_{t+1}|s^1_t), p^{a_t}(o_t|s^1_t), p^{a_t}(r_t|s^1_t)) \tag{1}$$

and encodes the structure and function that is common to all three Bayesian agents. As such, it encodes our assumptions about the knowledge and the cognitive mechanics of the decision makers (agents and participants) and reflects the scientific theory about human cognition that we formally test in the current study. More specifically, and as detailed on pp. 9 - 10 in the original manuscript, T is the decision horizon of the agent. The sets S, A, R, O are the outcome sets of the random variables that constitute the Bayesian agents and are necessarily specified in any probabilistic model, as random variables without outcome sets do not exist. The probability distributions $p(s^1_1), p(s^1_{t+1}|s^1_t), p^{a_t}(o_t|s^1_t)$ conjointly specify the generative probabilistic model that the agent is assumed to invert in the formation of its belief state. Without the specification of this model, the Bayesian evaluation of a belief-state is precluded. Together with the following comment by the reviewer, it seems that the reviewer assumes that merely "defining a belief state b_t" suffices to instantiate rational Bayesian agents. This is misled. A (Bayesian) belief state is a conditional probability distribution that is evaluated in the context of a probabilistic model that expresses the agent's probabilistic assumptions about its environment. Eqs. (19), (22), (23), and (24) all feature the agent belief state $p^{a_{1:t-1}}(s^1_t = 1|o_{t:t-1})$ which is deduced from $p(s^1_1), p(s^1_{t+1}|s^1_t), p^{a_t}(o_t|s^1_t), p^{a_t}(r_t|s^1_t)$ as documented in eq. (15) and Supplementary material S.4 and computationally implemented as documented in Supplementary material S.5. The evaluation of eqs. (19), (22), (23), (24) and their computational implementation thus rest critically on the definitions of the distributions $p(s^1_1), p(s^1_{t+1}|s^1_t)$, and $p^{a_t}(o_t|s^1_t)$. Finally, $p^{a_t}(r_t|s^1_t)$ is the unobserved state-conditional, action-dependent reward distribution that is central in the specification of action valence functions. In summary, all definitions in the general agent structure M_{Agent} are crucial for explaining the Bayesian agents in eqs. (19) - (24). In other words, without their definition, the Bayesian agents themselves could not be defined.

What's the point of defining $p^{a_{1:t-1}}(s^1_t|o_{1:t-1})$ in Eq. 15 when you will just redefine it as belief state b in

Eq. 16?

To address the reviewer's questions, we first emphasize that eq. (15) is not a definition (which in standard mathematical notation is denoted by $:=$), but an equality. Specifically, as explained on p. 11 lines 350 - 362, eq. (15) documents the form of the agent's probabilistic state representation (belief state) on trial t given the history of observations $o_{1:t-1}$ and actions $a_{1:t-1}$ under the joint distribution $p^{a_{1:t-1}}(s_t^1, o_{1:t-1})$ that is induced by the agent model documented on pp. 9-11. The derivation of the specific form of $p^{a_{1:t-1}}(s_t^1|o_{1:t-1})$ is detailed in Supplement S.4. and allows for the computational evaluation of the agent's belief in terms of stochastic matrix representations of the agent's prior, state-state transition, and action-dependent state-conditional observation distributions as implemented in the *belief_update* method of the agent class definition files *irb_agentobj_*.m*. In essence, eq. (15) constitutes the adaptation of a partially-observable Markov decision process filter equation to the particularities of the current model scenario (Åström, 1965; Cassandra et al., 1994; Kaelbling et al., 1998; Littman, 2009; Bäuerle and Rieder, 2011, see e.g.). Alternatively, eq. (15) may be viewed as a means to evaluate the filter distribution in an action-dependent hidden Markov model (see e.g., Rabiner, 1989; Murphy, 2012; Sarkka, 2013). The point of defining $b_t := p^{a_{1:t-1}}(s_t^1 = 1|o_{1:t-1})$ is to reduce the notational complexity in expressions that involve $p^{a_{1:t-1}}(s_t^1 = 1|o_{1:t-1})$. Finally, note that in eq. (16) $p^{a_{1:t-1}}(s_t^1|o_{1:t-1})$ is not redefined as b, but b_t refers to the filter probability that $s_t^1 = 1$. In other words, eq. (15) refers to a probability distribution while eq. (16) refers to a probability.

Major Issues

I found it difficult to keep track of the 4 different trial types. Perhaps replace "lucrativeness" with expected value, and lucrative/detrimental with rich vs. poor or pos/neg to make it more salient.

We thank the reviewer for this comment and take it as an opportunity to elaborate on the the naming of the experimental factors and their levels, which we have striven to choose in a careful manner and precisely define in the description of the Experimental design (pp. 3 - 4, lines 109 - 157). The advantage of using the terms "lucrativeness (lucrative vs. detrimental)" over the suggestions by the reviewer is that they directly evoke financial connotations and thus clearly demarcate this experimental manipulation from the "informativeness" manipulation. We argue that the term "expected value" is disadvantageous as soon as multiple random variables that model different concepts such as monetary rewards or information come into play, because every random variable has an expected value. Additionally, we would like to emphasize that there are only two, rather than four trial types, as also stated in our original submission on p.4:

"We refer to the trials with action choices between L $\wedge$ I and D $\wedge$ N actions as Trial Type I. Accordingly, on the other half of the trials, participants faced a choice between a lucrative and non-informative action (L $\wedge$ N) and a detrimental and informative (D $\wedge$ I) action. We refer to the trials with action choices between L $\wedge$ N and D $\wedge$ I actions as Trial Type II."

Replace intersection symbol with + or & for ease of reading.

We thank the reviewer for this comment and have replaced the intersection symbol with the more sensible *logical and* symbol $\wedge$.

Roman numerals for the four trial types are introduced in the text, but never used in the figures. Perhaps a more informative naming of the trial types and the pairwise differences between them could be introduced to make the result plots easier to understand later.

We thank the reviewer for this comment and have included the Trial Type identifiers in the revised version of Figure 1, as shown below in response to another comment of the reviewer on Figure 1.

It's quite clear that the paper is focused on the models rather than behavior, but page 18 seems far too late to introduce the first behavioral results plot. Perhaps first present the behavioral results before describing the models and presenting the modeling results. This may somewhat help the reader build intuitions about what the models are capturing

We thank the reviewer for this comment, but consider the order of content presentation in our manuscript an editorial affair. In our presentation, we followed the classical article outline of

(1) Introduction

(2) Methods

(3) Results

(4) Discussion

As the computational models are a mean to condense and interpret the participant's choice data sets, they are clearly of methodological nature and hence part of (2) Methods. The descriptive statistics of the participant's choice data sets, on the other hand, are an empirical result and hence part of (3) Results. If advised by the editor, we would be happy to restructure the article in the form

(1) Introduction

(2) Experimental and descriptive statistical methods

(3) Descriptive results

(4) Mathematical and computational methods

(5) Model validation and evaluation results

(6) Discussion

As this outline would strongly deviate from the standard article format in Computational Brain and Behavior, we leave it to the editor to decide whether this outline is preferable.

Finally, with respect to the reviewer's statement that "the paper is focused on models rather than behaviour", we argue for the opposite and emphasize that the agent models are devised as a means to interpret the participants' trial-by-trial choice data. The observed behavioral data is thus at the heart of the current study and used to arbitrate between different cognitive computational theories used for their explanation.

I also was confused why the state-state transition distribution was fixed to the true values rather than inferred, using something like a beta-binomial distribution for example.

We thank the reviewer for this comment. As we stated on p.9 of the original version of our manuscript, the state-state transition probability encoded by the agent models was fixed to the true value, because the participants were explicitly informed about it:

"Crucially, participants were informed that a reversal would happen 1-4 times in a run, but were not informed about the approximate number of trials without a reversal. Therefore, we equipped the agent with a constant reversal probability of 0.0375, which reflects the true reversal frequency in a run (there were 3 reversals across the 80 trials)."

The participants thus had explicit knowledge of the state-state transition probability, and the same assumption was thus made for the agent models.

The softmax choice policy (Eq. 26) is part of the models, and not part of the model evaluation. This is making specific theoretical commitments and should be included in the previous section

We thank the reviewer for this comment. As the reviewer raises this point on multiple occasions, we address it in a unified manner below.

The BIC definition in Eq. 29 is wrong. There is a sign inversion here. Also, if you are dividing through by 2, then it's not the exact BIC but rather a measure proportion BIC but 1/2 the magnitude.

We thank the reviewer for raising the important issue of the definition of the BIC, but we disagree with the reviewer's comment. As we wrote on p. 16, lines 479-480: "(...) we first evaluated all agent model- and participant-specific Bayesian Information Criterion scores according to (Schwarz et al., 1978)" and on p. 16, lines 483-486 "The BIC scores of all agents and participants were then subjected to a random-effects Bayesian model selection procedure (...) (Rigoux et al., 2014)." (we apologize for the misformatting of the reference to Schwarz (1978) in the original version of our manuscript). Before we delve into the reasoning for using the form

$$\text{BIC} := \ell(\hat{\theta}) - \frac{k}{2} \ln N \tag{2}$$

as suggested by Schwarz (1978) (rather than a different form as implied by the reviewer's comment), we first review some necessary background on the BIC following Held (2008).

Let

$$p(y, \theta) = p(y|\theta)p(\theta) \tag{3}$$

denote a probabilistic model, i.e. a joint distribution of data y and parameters θ. The fundamental motivation for evaluating BIC scores is to obtain an approximation to the log marginal likelihood

$$\ln p(y) = \ln \int p(y, \theta)\, d\theta \tag{4}$$

of a probabilistic model, which constitutes a core component in Bayesian model comparison (e.g. Kass and Raftery, 1995). The BIC approximation of the log marginal likelihood rests on Laplace's integral approximation method, which we review for the case of a multivariate real-valued function first.

Definition 1 (Laplace's integral approximation method). Let $f : \mathbb{R}^k \to \mathbb{R}$ denote a twice-differentiable convex multivariate real-valued function with a minimum at $\tilde{x} \in \mathbb{R}^k$. Then *Laplace's integral approximation method* corresponds to approximating the integral

$$I_N := \int_{\mathbb{R}^k} \exp(-Nf(x))\, dx \tag{5}$$

for $N \in \mathbb{N}$ according to

$$I_n \approx \left(\frac{2\pi}{N}\right)^{\frac{k}{2}} |H^f(\tilde{x})|^{-\frac{1}{2}} \exp(-Nf(\tilde{x})), \tag{6}$$

where $|H^f(\tilde{x})|$ denotes the determinant of the Hessian of f evaluated at $\tilde{x}$. •

For the motivation and derivation of Laplace's integral approximation method, see e.g. Lauritzen (2009). Using Laplace's integral approximation method it can readily be shown that an approximation of the log marginal likelihood $\ln p(y)$ is afforded by

$$\ln p(y) \approx \frac{k}{2} \ln 2\pi - \frac{k}{2} \ln N - \frac{1}{2} \ln |H^f(\hat{\theta})| + \ln p(y|\hat{\theta}) + \ln p(\hat{\theta}), \tag{7}$$

where $\hat{\theta}$ denotes a maximum of the function

$$\ell : \mathbb{R}^k \to \mathbb{R}, \theta \mapsto \ell(\theta) := \ln p(y|\theta). \tag{8}$$

Proof. Define

$$f : \mathbb{R}^k \to \mathbb{R}, \theta \mapsto f(\theta) := -\frac{1}{N} \left(\ln p(y|\theta) + \ln p(\theta)\right), \tag{9}$$

and let $\tilde{\theta}$ denote a minimum location of f. Then we have with the Laplace integral approximation of eq. (6)

$$\begin{aligned}
\ln p(y) &= \ln \int p(y|\theta)p(\theta)\, d\theta \\
&= \ln \int \exp\left(\ln\left(p(y|\theta)p(\theta)\right)\right)\, d\theta \\
&= \ln \int \exp\left(\ln p(y|\theta) + \ln p(\theta)\right)\, d\theta \\
&= \ln \int \exp\left(-N\left(-\frac{1}{n}\left(\ln p(y|\theta) + \ln p(\theta)\right)\right)\right)\, d\theta \\
&= \ln \int \exp\left(-Nf(\theta)\right)\, d\theta \\
&\approx \ln \left(\left(\frac{2\pi}{N}\right)^{\frac{k}{2}} |H^f(\tilde{\theta})|^{-\frac{1}{2}} \exp(-Nf(\tilde{\theta}))\right) \\
&= \frac{k}{2} \ln 2\pi - \frac{k}{2} \ln N - \frac{1}{2} \ln |H^f(\tilde{\theta})| + \ln p(y|\tilde{\theta}) + \ln p(\tilde{\theta})
\end{aligned} \tag{10}$$

□

The BIC score as stated in eq. (29) of our manuscript then results from (c.f. Held, 2008)

- neglecting the constant terms $\frac{k}{2}\ln 2\pi$ and $\ln p(\hat{\theta})$ and
- neglecting the term $-\frac{1}{2}\ln|H^f(\tilde{\theta})|$, which for $n \to \infty$ is bounded from above.

With $\ln p(y|\hat{\theta}) = \ell(\hat{\theta})$, it then follows that

$$\ln p(y) \approx \ell(\hat{\theta}) - \frac{k}{2}\ln N =: \text{BIC}. \tag{11}$$

The motivation for using the BIC score formulation as defined in eq. (29) of our manuscript and derived in eq. (11) herein is as follows. First, the group-level model comparison approach proposed by Rigoux et al. (2014) which we use for the evaluation of protected exceedance probabilities (PEPs, quantifying the group-level probability that any one model is more likely than the others above and beyond chance) requires (positive) participant-specific log model evidence approximations, rather than negative participant-specific log model evidence approximations. Second, while we are aware that different versions of the BIC exist in the literature, we explicitly referred the reader to Schwarz (1978). Therein, we read on pp. 461-462

"An extension of the maximum likelihood principle is suggested by Akaike [1] for the (...) problem of choosing among different models with different numbers of parameters. His suggestion amounts to maximizing the likelihood function separately for each model j, obtaining, say, $M_j(X_1, ..., X_n)$, and then choosing the model for which $\ln M_j(X_1, ..., X_n) - k_j$ is largest, where k_j is the dimension of the model. We present an alternative approach to the problem. (...) We shall arrive at the following procedure:

Choose the model for which $\ln M_j(X_1, ..., X_n) - \frac{1}{2}k_j \ln n$ is largest.

The validity of this procedure as a large-sample version of Bayes procedures will be established here for the case of independent, identically distributed observations, and linear models."

Our formulation of the BIC in eq. (29) is thus in line with the reference provided in our original submission. We are afraid that we do not see and advantage of the suggested alternative formulation and, given that a reference that we could consult in this regard was not provided, we opted for retaining our initial formulation.

Also, since the models are fit to the full data, the pseudo-R$\hat{2}$ in Eq. 30 would be better described using the BICs than the unpenalized log likelihoods

We thank the reviewer for this comment. As described above, our model comparison procedure and model comparison results rest on the evaluation of BIC scores as approximations to the log model evidence at the participant level and the evaluation of protected model exceedance probabilities at the group level (Rigoux et al., 2014). We only included a pseudo-r^2 measure for those readers, for whom such a measure is more familiar and reported the pseudo-r^2 result in a single sentence and the end of the Model comparison results section. Clearly, in addition to the BIC/PEP approach, there exists a plethora of other scalar measures of model fit (e.g., Long, 1997, Section 4.3), none of which shows clear conceptual benefits over the BIC/PEP score approach employed here. For internal coherence, we thus opted to remove any reference to the pseudo-r^2 measure from the revised manuscript.

The statistics for the behavioral results are largely incomplete. An ANOVA comparing the 2 x 2 manipulations of payoff and informativeness would be greatly appreciated.

We thank the reviewer for raising the issue of analyzing the data by means of a 2 x 2 ANOVA. However, while the experimental design indeed manipulated the action's lucrativeness (lucrative, detrimental) and informativeness (informative, non-informative) in a 2 x 2 fashion, participant actions were recorded only on two trial types, contrasting (1) a lucrative and informative action with a detrimental and non-informative action (Trial Type I in the original version of our manuscript, cf. p. 4, lines 147 - 154) and (2) a lucrative and non-informative action with a detrimental and informative action (Trial Type II in the original version

of our manuscript, cf. p.4, lines 147 - 154). Participant-specific choice rates across the four action conditions are thus not independent: for example, if a participant selects a lucrative and informative action, they always decide against a detrimental and informative action. A 2 x 2 ANOVA experimental design and statistical evaluation rests on the independent acquisition of data for all four design cells, which is clearly not the case for the current experiment. A 2 x 2 ANOVA thus simply does not apply for the current behavioural data set. Furthermore, we note that in the current study, we are interested in the statistical modeling of trial-by-trial choice behaviour by cognitive agents rather than the linear modeling of descriptive summary statistics. Yet, by formatting the data according to the Brain Imaging Data Structure (Gorgolewski et al., 2016) and making it openly available under `https://osf.io/vdmah/`, we believe to have laid the prerequisite for readers to further scrutinize the data, and, if of interest, do this in terms of linear modeling of the descriptive summary statistics.

The sole t-statistic is described quite strangely, and I would advise sticking to APA guidelines.

We thank the reviewer for this comment. In our submission, we wrote

> To statistically corroborate this finding, we conducted a two-sided paired sample t-test across participants. This yielded a test statistic of $t(23) = 11.55$ with an associated p-value smaller than 0.001.

Given that the guidelines of the APA have not been developed for interdisciplinary research, in our opinion, they have their limits when it comes to reporting exploratory computational research as presented in our manuscript. However, following the reviewer's suggestion, we reviewed the APA guideline as available from `https://apastyle.apa.org/jars/quant-table-1.pdf`. We read

> Provide information detailing the statistical and data-analytic methods used, including (...) inferential statistics, including results of all inferential tests conducted, including exact p values if null hypothesis significance testing (NHST) methods were used, and reporting the minimally sufficient set of statistics (e.g., dfs, mean square [MS] effect, MS error) needed to construct the tests.

In our description of the statistical evaluation of the rate of lucrative choices if the lucrative action was informative vs. that the lucrative action was non-informative, we included the type of statistical test ("two-sided paired sample t-test"), the value of the test statistic and the degrees of freedom of the test ("$t(23) = 11.55$"), as well its its associated exact p-value, to the commonly accepted level of detail ("smaller than 0.001"). We thus believe that we provide reader with all necessary information about this test and indeed follow the APA guidelines.

This comparison of performance across conditions should ideally be visualized in the behavioral results plot (fig. 3).

We thank the reviewer for this comment and have included the average choice rates in Figure 3 as shown below in the context of another comment of the reviewer on this figure.

The discussion says that "we here do not interpret the softmax operation as random exploration", but I wonder if the softmax temperature is related to performance or is correlated with the lambda weighting parameter. Do directed and random exploration trade off with one another? How else should we interpret the modeling results if we cannot interpret the parameter estimates?

We thank the reviewer for this comment. As the reviewer raises this point on multiple occasions, we address it in a unified manner below.

Also, what does lambda tell us about participant behavior? Is there an intermediate balance between exploitation vs. exploration that is optimal in the task? How close were participants to this value?

The reviewer highlights the intriguing issue of an optimal choice policy. However, the derivation of an optimal choice policy by means of theory of partially-observable Markov decision processes is beyond the scope of the current work. By explicitly specifying the task and agent models, as well as sharing the behavioral data, we believe to have laid the foundations for such an undertaking by us or interested readers. However, given the extent of the current manuscript, we have to leave its implementation to future work. We address this point in a new paragraph of the Discussion section:

At last, some limitations of our study along with some suggestions for future research must be acknowledged. First, as for any scientific study, inferences can only be made with respect to the model space, which *per se* is incomplete. Thus, although we provide evidence for a strategy resembling a belief state-based agent seeking to maximize the convex combination of its expected Bayesian surprise and its expected reward, it is very well possible that an agent not included in our model space can better account for the behavioral data. Based on our experimental documentation and the open availability of the data, future research may broaden the model space and could also consider, for example, probabilistic variants of the win-stay-lose-switch agents (Worthy et al., 2012) or agents with non-adaptive learning rates (Wiering and Otterlo, 2012; Rescorla and Wagner, 1972), besides the already mentioned random exploration agents in a probabilistic filtering framework (Ostwald et al., 2014a). Second, on a related note, we here do not derive an optimal agent in the sense of partially observable Markov decision process theory (Bertsekas, 2000; Puterman, 2005; Bäuerle and Rieder, 2011). A model space that comprises an agent implementing an optimal choice strategy would afford the analysis of the participants' behavior from a normative perspective and is thus an interesting endeavor for future research. To facilitate such an undertaking, we have provided a detailed documentation of the task and agents developed thus far. Third, given the modest sample size of our study, the behavioral data reported here are best capitalized on in an exploratory fashion. Together with the implemented open research measures (Ritchie, 2020), we believe to have laid the foundations for reproducing our study and build upon our work.

Minor issues

Formatting. Citations on line 21 are formatted incorrectly. Should not have parentheses inside parentheses.

We thank the reviewer for pointing out this formatting error and have removed the additional parentheses and additional reference formatting errors throughout the manuscript.

Use of Ibid on line 27 isn't APA.

We thank the reviewer for this comment. Indeed, the use of Ibid is not discussed in the 7th Edition of the APA Publication Manual. However, in the context of respective sentence we consider it useful and Computational Brain and Behaviour does not mandate any form of adherence to the suggestions of the APA Publication Manual. We thus argue for retaining the use of Ibid on line 27.

Introduction. Line 39-40 is not a question.

We thank the reviewer for pointing out this imprecision and have rewritten the sentence as follows:

The pure exploration paradigm thus raises the problem of how to strike a balance between the exploration costs preceding the final trial and the potential final trial reward (Ostwald et al., 2015a).

ADHD scores. The description for participants initially made me a bit suspicious, but when I consulted with an ADHD expert, I was told that the range of Conners values are all below even what is considered sub-clinical levels (<70). You should consider re-writing the Participants section to create less of a red-herring for the reader.

We thank the reviewer for engaging with the description of the study sample in detail. From our perspective, the Participants section concisely expresses the participant sampling procedure, including the fact that while the initial sampling had the study of ADHD in mind, the final sample was treated as a healthy sample with respect to ADHD. As the reviewer correctly notes, the CAARS-S:S T-scores are below sub-clinical levels. From the perspective of this test, the participants can thus be considered healthy, which we state in the original submission as "We (...) treated the group of participants as a healthy sample and did not conduct analyses to relate ADHD symptoms to task behaviour."

Task description. Use plain language for more effective communication. "Depending on the laterality of the shape chosen" could just "depending on the side". There are many other cases of obtuse language that hinders interpretability.

We thank the reviewer for this comment and have rewritten the respective statement as

Depending on the side of the shape chosen, the action was also either informative and the returned reward was revealed to the participant, or it was non-informative and the returned reward was not revealed to the participant.

It's not necessarily to define what a maximum likelihood is. Rather, a clear description of the free

parameters in each model would be more useful here.

We thank the reviewer for this comment. As the reviewer raises similar points with regards to the softmax function on multiple occasions, we address it in a unified manner below.

Table 1 should either be moved to the supplement or visualized as part of Fig. 1. It's not very accessible

We thank the reviewer for this suggestion, but point out, that Table 1 simply states the probability mass functions that characterize the task model and thus precisely defines the probabilistic environment that the participants and agent models interact with. We thus argue that Table 1 conveys essential information in the documentation of the experiment and modeling initiative and forms an integral part of the Methods section for any reader who would like to reproduce or build upon our work. In addition, by using symbols that are defined in the main text and described in the table caption, we believe to have rendered the table accessible. From our perspective, a bar graph visualization of the probability mass functions would not provide further insights.

Figs. 2, 3, and 6 are missing y-axis labels, making it hard to interpret the results.

We thank the reviewer for this comment. In the revised versions of Figures 2,3, and 6, we include y-axis labels on all subpanels, whenever appropriate.

FIGURE 1

Fig. 1. What was the point of using the roman numeral type numbers if they're not illustrated in Fig. 1a? Alternatively, find more informative type names and prominently display them here. There is also no point in displaying the Informative vs non-informative reward rates, in panel d, since that is an orthogonal manipulation

We thank the reviewer for this comment. In the revised version of Figure 1, we now include the trial type denominations and omit the display of the observed reward rates as shown below.

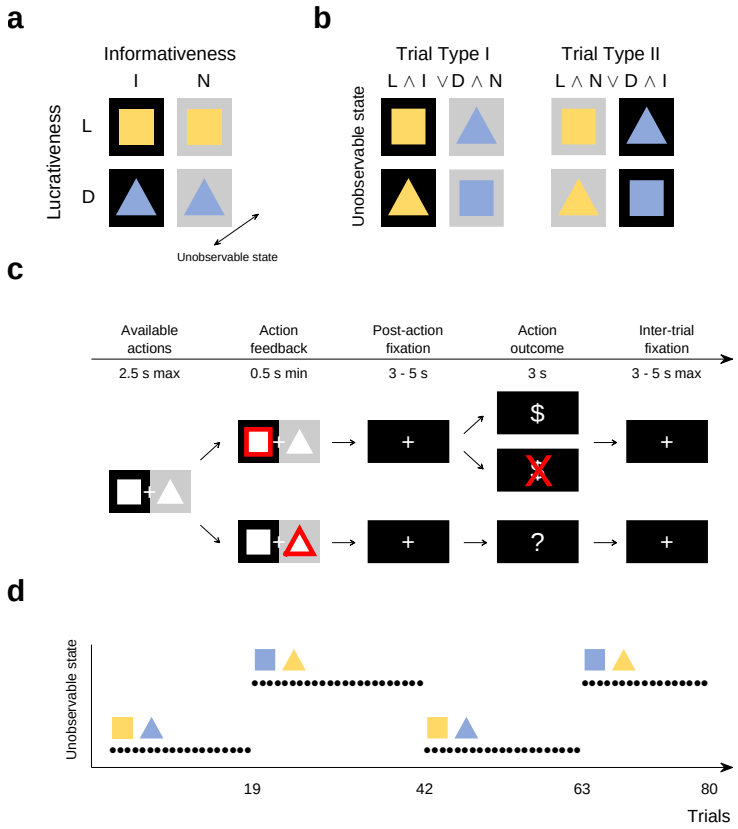

Figure 1: Information-selective symmetric reversal bandit task. a Experimental design. Possible actions differed in lucrativeness (lucrative (L) or detrimental (D)), as well as in informativeness (informative (I) or non-informative (N)). The former experimental factor was associated with selectable shapes (square and triangle), while the latter experimental factor was associated with black and grey screen sides. In addition, and unbeknownst to the participants, the lucrativeness of the shapes reversed at random trials throughout the experiment, corresponding to the unobservable task state: for some time, the square may represent the lucrative action, indicated here by its yellow color (and the triangle, accordingly, represent the detrimental action, indicated here by its blue color), but this could reverse. **b** On a given trial, participants faced a choice between either a lucrative and informative vs. a detrimental and non-informative action (Trial Type I; L ∧ I or D ∧ N) or a lucrative and non-informative vs. a detrimental and informative action (Trial Type II; L ∧ N or D ∧ I). **c** Trial design. Participants could indicate their choice within 2.5 seconds of the choice options onset. If they chose the shape on the black side, the returned reward was revealed (top). If they chose the shape on the grey side, the returned reward was not revealed (bottom). Note that the current lucrativeness of the shapes was not revealed to the participants, hence the white shape color. **d** Run design. Every 17 to 23 trials the reward probabilities associated with the shapes reversed. Here, the reversal times of the first run are shown. For trials 1 to 19 the square was lucrative, indicated by its yellow color, and the triangle was detrimental, indicated by its blue color. This reversed on trial 20 at which choosing the triangle became the lucrative action and choosing the square became the detrimental action. For the generation of this figure, please see *figure_1.m*.

FIGURE 2

Fig. 2 uses different colors in the top left corner, but it's not clear what this difference refers to. I would rather see the left and right panels overlaid and the use of a single color gradient corresponding to different values of the a parameter. I'm also not sure how informative this plot is, since it doesn't offer any validation from participant data, since you don't have access to beliefs. Perhaps introducing model predictions here as a function of some observable feature of participant behavior would be more informative.

We thank the reviewer for this comment. In fact, Figure 2 uses the same agent color scheme as used for all figures of the manuscript, but we agree that this was rather implicit. We now included an explicit reference to the color scheme in the caption of Figure 2:

Note that the colors used for the graphical depiction of the agents' action valence functions correspond to the agent model color scheme used for model recovery and model comparison in all Figures below.

We also revised the general layout of the figure to make it more accessible, as shown below.

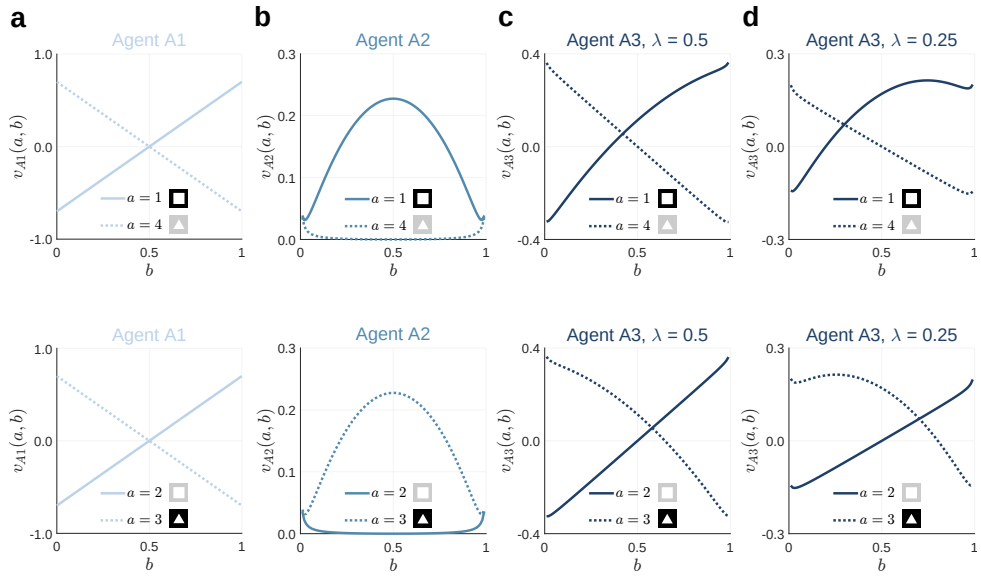

Figure 2: Bayesian agents' action valence functions. **a** Action valence functions of agent A1 for the available action set $A_1 := \{1, 4\}$ (upper panel) and the available action set $A_2 := \{2, 3\}$ (lower panel). Agent A1 allocates action valences based on the belief state-weighted expected reward. As the expected rewards for choosing the lucrative or detrimental shape are constant, the more extreme the agent's belief that a given shape is lucrative the higher the valence it allocates to choosing the corresponding shape and the lower the valence it allocates to choosing the other shape. The valences of A1 do not depend on the actions' informativeness, which reverses between available action sets A_1 and A_2 and therefore the two panels are identical. **b** Action valence functions of agent A2 for the available action set $A_1 := \{1, 4\}$ (upper panel) and the available action set $A_2 := \{2, 3\}$ (lower panel). Agent A2 allocates action valences based on the expected Bayesian surprise, which is higher for the informative action than for the non-informative action, thus the graphs on the upper and lower panels flip. The lower the agent's uncertainty about the lucrative shape, the larger the benefit of the informative action. **c, d** Action valence functions of agent A3 for the available action set $A_1 := \{1, 4\}$ (upper panel) and the available action set $A_2 := \{2, 3\}$ (lower panel) for $\lambda = 0.5$ (**c**) and $\lambda = 0.25$ **d** Agent A3 allocates action valences based on the convex combination A1 and A2 action valences. The higher the value of λ the more the valences of A3 resemble the valences of A1 and correspondingly, the lower the value of λ the more the valences of A3 resemble the valences of A2. Note that the colors used for the graphical depiction of the agents' action valence functions correspond to the agent model color scheme used for model recovery and model comparison in all Figures below. For implementational details, please see *figure_2.m*.

With respect to the informativeness of this plot, we note that the figure visualizes the action valence functions of the Bayesian agent models A1, A2, and A3. The figure thus subserves the understanding of the agent's valence functions formulated in eqs. (19), (22), and (23). In other words, the aim of the figure is not to convey any relationship between the agent model predictions and human behaviour (this is the aim of Figure 6), but to demonstrate the mathematical mechanics of the implemented agent models.

FIGURE 3

Fig. 3. Include the performance between conditions as reported at the top of the Descriptive statistics section. The color/shape manipulation in panel a is very difficult to understand. Use more differentiable colors and/or distinct shapes (e.g., triangle vs. circle). Condition specific learning curves could also be a more informative alternative.

We thank the reviewer for this comment and have included the action-specific choice rates in the revised version of Figure 3a. We also revised the color/shape manipulation for the visualization of the four different action choice rates as laid out in the context of Figure 1. Specifically, we now contrast lucrative and detrimental actions by yellow and blue and informative and non-informative actions by black and grey, respectively. This color manipulation is perceptually salient and identical across the visualization of all descriptive statistics.

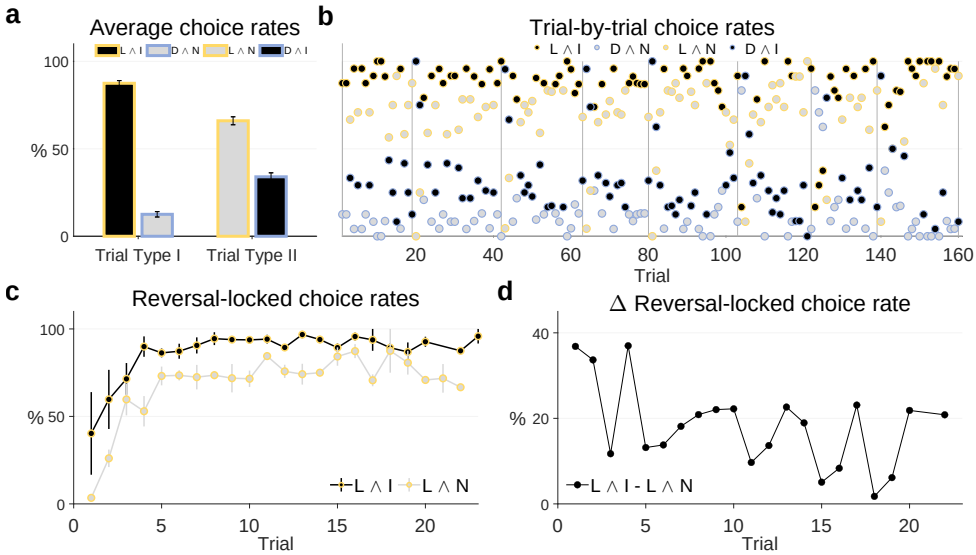

Figure 3: Participant choice rates. a Average choice rates across participants and task dynamics for Trial Type I and Trial Type II. On Trial Type I, participants faced a choice between a lucrative and informative (L ∧ I) and a detrimental and non-informative (D ∧ N) action. On average, participants preferred the L ∧ I action. On Trial Type II, participants faced a choice between a lucrative and non-informative (L ∧ N) and a detrimental and informative (D ∧ I) action. On average, participants preferred the L ∧ N action, but to a lesser degree than the L ∧ I action on Trial Type I. **b** Group trial-by-trial L ∧ I, D ∧ N, L ∧ N, and D ∧ I action choice rates. The vertical lines represent the last trial before a reward rate reversal. The vertical lines at $t = 80$ and $t = 160$ mark the end of the first and second run, respectively. Note that given the experimental design, the action choice rates on Trial Type I (L ∧ I vs. D ∧ N) and on Trial Type II (L ∧ N vs. D ∧ I) are complementary and add up to 100%. On the majority of trials the lucrative action choice rates L ∧ I and L ∧ N prevail over the detrimental action choice rates D ∧ I and D ∧ N. This effect is more pronounced for the L ∧ I choice rate than for the L ∧ N choice rate. **c** Average reversal-locked group trial-by-trial action choice rates for L ∧ I and L ∧ N actions. Both action choice rates increase over post-reversal trials. The error bars depict the SEM over reversal blocks. **d** Average group reversal-locked L ∧ I and L ∧ N choice rate difference. The difference decreased between the first trials after and the last trials before a reversal. For implementational details, please see *figure_3.m*.

With respect to the “condition-specific learning curves” alternative suggestion, we note that Figure 3c continues to depict exactly those: in the original and revised versions of our manuscript, we write (ll. 554 - 564 of the original version)

Notably, as shown in Figure 3c, both the average L ∧ I action choice rate as well as the average L ∧ N action choice rate exhibit an overall increase between two reward rate reversals, indicating that participants were able to gradually resolve their uncertainty about the currently lucrative shape. Moreover, as shown in Figure 3c, although the average L ∧ I action choice rate was larger than the average L ∧ N action choice rate on virtually all trials between two reversals, their difference decreased slightly between the first trial after and the last trial before a reward rate reversal (Figure 3d). This suggests that with decreasing uncertainty about the currently lucrative shape participants placed less valence on the actions’ informativeness.

The conditions of our study are the available actions L ∧ I (complementary to D ∧ N) and L ∧ N (complementary to D ∧ I). Figure 3c depicts the reversal-locked (= post-reversal) choice rate evolution for these conditions (where, as in Figure 3b, the complementary D ∧ N and D ∧ I choice rates can be inferred, because the respective choice rates add to 100 %). Figure 3c thus continues to depict “condition-specific learning curves”. In addition, Figure 3d continues to depict the differential effect of the informativeness condition on the learning of which shape currently is lucrative: there is a slight decrease in the advantage of choosing the L ∧ I action over the L ∧ N as learning proceeds, implying that participants value information less the more certain they are about the currently lucrative shape.

FIGURE 4

Fig. 4. Perhaps use an alternative qualitative color scheme, because it is very difficult to identify which model is best.

We thank the reviewer for pointing out this issue. Accordingly, we have made the agent color scheme more perceptually salient by using green colors for the two control models C1 and C2 as shown below.

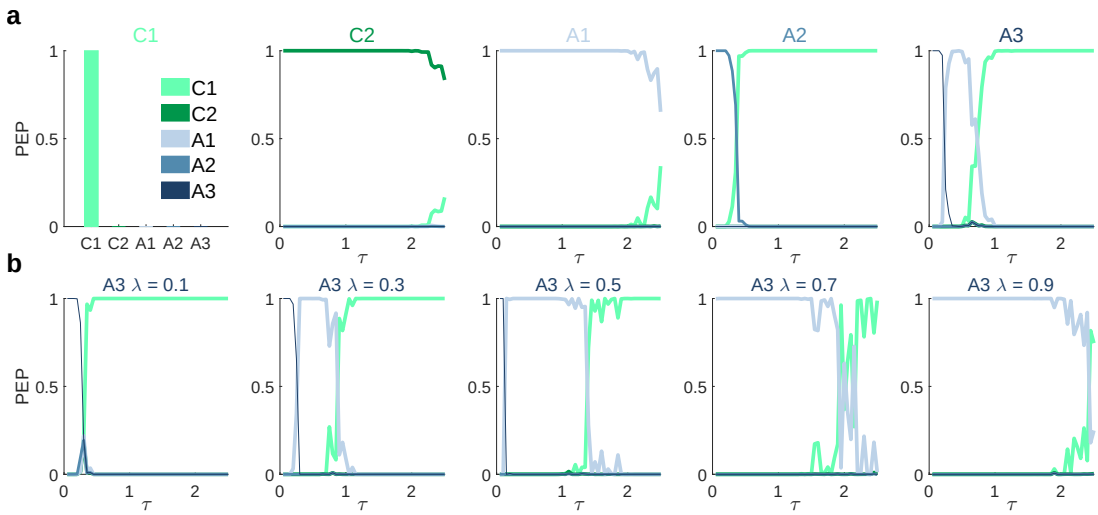

Figure 4: Model recovery results. **a** Model recovery results. Each subplot pertains to a specific data generating model (C1, C2, A1, A2, and A3) and shows the protected exceedance probabilities (PEPs) of all data analysis models in the model space evaluated in light of data sets sampled from the respective data generating model. For data generated based on the cognitive null model C1, the PEP was maximal for C1. For data generated based on agents C2, A1, A2, and A3, the PEPs depend on the value of the post-decision noise parameter τ. Agents C2 and A1 are recoverable up to high levels of post-decision noise. Agents A2 and A3 (for $\lambda = 0.25$) are recoverable only for low levels of post-decision noise. **b** Model recovery results for agent A3 for different values of the action valence weighting parameter λ. Agent A3 is recoverable up to medium values of λ and for low levels of post-decision noise. For implementational details, please see *figure_4.m*.

FIGURE 6

Fig. 6. BIC is inverted as discussed above. The behavioural validation in panel b is hard to compare with the behavioral plot in Figure 3. Maybe find a way to combine participant vs. model performance on a single plot.

We thank the reviewer for this comment. As discussed above, the BIC - as computed according to (Schwarz et al., 1978) - is not inverted. To address the reviewer's comment with respect to the difficulty to compare the reversal-locked choice rates of the simulated behavioural data (Figure 6b) with the reversal-locked choice rates of the observed behavioural data (Figure 3b, 3c), we now overlay the simulated reversal-locked choice rates on the empirically observed reversal-locked choice rates in the left and right panels of Figure 6b. As shown in the revised version of Figure 6 below, thet data are in good agreement.

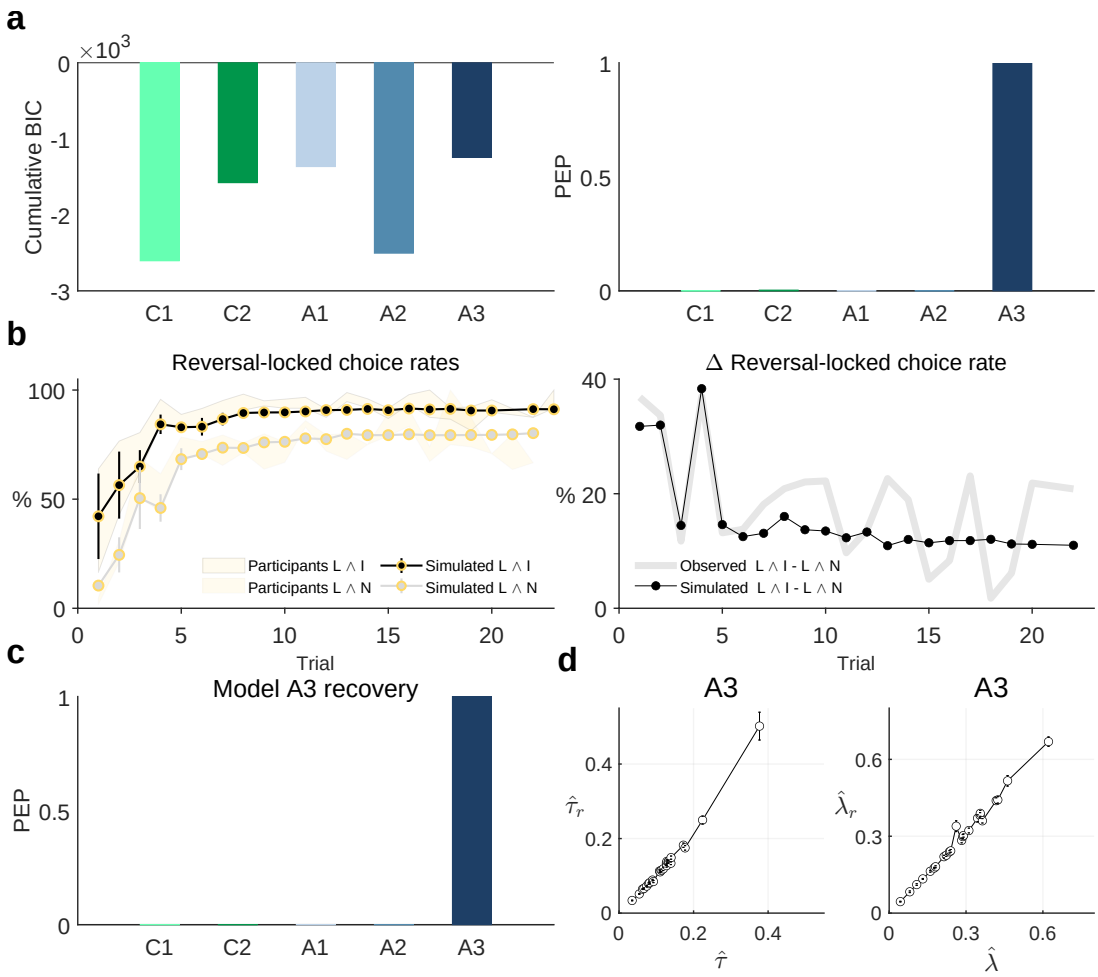

Figure 5: Computational modelling results. **a** Model comparison results. Both the group cumulative Bayesian information criterion scores (cumulative BIC, left panel) and the protected exceedance probabilities (right panel) were maximal for the agent model A3, indicating that this model explained participants' choice data the best. **b** Model A3 behavioural validation. Average reversal-locked group trial-by-trial L ∧ I action and L ∧ N action choice rates (left panel) and their difference (right panel) based on synthetic data sets generated using agent A3 and the participants' parameter estimates. The patterns closely resemble those observed in the participants' data (cf. Figures 3c, 3d), re-visualized here for convenience. **c** Model A3 recovery based on the participants' data parameter estimates. The protected exceedance probability based on re-analysis of simulated data from agent model A3 with parameters set to the participants' parameter estimates is maximal for agent model A3. **d** Parameter recovery results based on data generated with agent model A3 and the participants' data-based parameter estimates. The panels depict the simulation-averaged participant-specific recovered parameter estimates $\hat{\tau}_r$ and $\hat{\lambda}_r$ and their associated SEMs over simulations as a function of the participant-specific parameter estimates $\hat{\tau}$ and $\hat{\lambda}$. Both the participants' post-decision noise parameter estimates $\hat{\tau}$ as well as the participants' weighting parameter estimates $\hat{\lambda}$ are recoverable. For implementational details, please see *figure_6.m*.

Also, how are the different parameter estimates of A3 related to performance or behavior?

We thank the reviewer for this suggestion. The parameters of agent model A3 are adapted in such a manner that they optimally explain the participants' observed behavior from the perspective of maximum likelihood estimation. The lower the value of $\hat{\tau}$, the more closely does the participant behavior follow the respective choice strategy. The smaller the value of $\hat{\lambda}$, the more closely the participant behavior resembles the choice strategy of agent model A2.

win-stay-LOSE-shift instead of LOOSE

We thank the reviewer for pointing out this typo and have corrected it.

Almost all the abstract is spent describing the methods but only a single sentence is given to the results, with very little interpretation or take home message. It's not very informative to state that you use maximum likelihood to evaluate your data. I would rather like to see a more elaboration of your take home message. What do your model results tell us about human behavior we didn't already know? The task may be novel, but what is gained from this novel design? What new insight about human cognition is gained?

We thank the reviewer for drawing attention to our abstract. Given the scope of our study, we believe that the take-home message of our work is precisely, concisely, and modestly expressed in the last sentence of the Abstract:

In brief, we provide quantitative evidence that participants employ a belief state-based hybrid explorative-exploitative strategy on the information-selective symmetric reversal bandit task, lending further support to the finding that humans are guided by their subjective uncertainty when solving exploration-exploitation dilemmas.

I don't think it is justified to call tau "post"-decision noise, since you do not measure beliefs about value and thus have no evidence that this is noise due to how values are mapped onto actions. Temperature would be a better name

We thank the reviewer for raising attention to the issue of our use of the softmax operation for maximum likelihood estimation on multiple occasions, which we address in a unified manner here. From above, we recall the following additional reviewer comments in this regard:

- *The softmax choice policy (Eq. 26) is part of the models, and not part of the model evaluation. This is making specific theoretical commitments and should be included in the previous section.*
- *The discussion says that "we here do not interpret the softmax operation as random exploration", but I wonder if the softmax temperature is related to performance or is correlated with the lambda weighting parameter. Do directed and random exploration trade off with one another? How else should we interpret the modeling results if we cannot interpret the parameter estimates?*
- *It's not necessarily to define what a maximum likelihood is. Rather, a clear description of the free parameters in each model would be more useful here.*

We reason that the comments conflate the concepts of softmax decision rules (e.g., Sutton and Barto, 1998, Section 2.3) with the statistical embedding of deterministic agent models. Following the distinction of deterministic and randomized decision rules in Markov decision process theory (e.g., Puterman, 2005, Section 2.1.4) this issue is best addressed by first differentiating the notions of *deterministic* and *random agents*. By a *deterministic agent*, we understand a function that maps a history of observations and actions onto an action. Agents A1, A2, and A3 are deterministic: given a specific sequence of observations and rewards, they always emit the same action. Also note that the agent's belief state is deterministic, despite the fact that it is a probabilistic representation of the task state: given identical priors and observations, the identical probability for the state being 1 ensues. By a *random agent*, we understand a probability distribution that is parameterized by a history of observations, rewards and actions, and from which a (random) action can be realized. The agents in the current study are not random. Softmax decision rules are typical examples for random agents. As documented in eq. (18) of the original version of our manuscript, however, the decision rules in the current study take the form

$$d : \mathbb{R} \times [0,1] \to A_{s^2}, v(\cdot, b) \mapsto d(v(\cdot, b)) := \arg\max_{a \in A_{s^2}} v(a, b). \tag{12}$$

and are hence not of the softmax type.

What then is the role of the softmax operation in the definition of the log likelihood function in eq. (28) of the original submission? As clearly stated in the original submission, the softmax operation here allows for embedding the (deterministic) agent models in a statistical inference framework. An analogy to the well-known general linear model (GLM) may help to clarify this point. Given an $n \times p$ design matrix X, a $p \times 1$ effect parameter β, and positive σ^2, the GLM describes the distribution of $n \times 1$ observable data vectors as

$$y \sim N(X\beta, \sigma^2 I_n) \Leftrightarrow y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n) \tag{13}$$

where $N(\mu, \Sigma)$ denotes the multivariate normal distribution with expectation parameter μ and covariance matrix parameter Σ. Like the data analysis models specified in the Section "Model evaluation" of our original submission, the GLM comprises a deterministic component that conforms to the prediction of the model ($X\beta$) and a random component (ε) which models stochastic deviations from this prediction. The degree of deviation of the observed data y from the deterministic prediction of the model is governed by the variance parameter σ^2. The analogous concepts are at work in the definitions of $p(a_t = a|a_{1:t-1}, o_{1:t-1})$ in eq. (26) and the log likelihood function ℓ in eq. (28): the deterministic data prediction of each agent model is given by $\arg\max_{a \in A_{s^2}} v(a, b)$, whereas the random deviation from this prediction is modeled by emdedding $v(a, b)$ in the softmax operation. The degree to which the data prediction of the data analysis model corresponds to the agent model prediction is governed by the parameter τ: for $\tau \to 0$, the behavioral data prediction is identical to the agent model prediction (corresponding to the case of $\sigma^2 \to 0$ in the GLM scenario), while for $\tau \to \infty$, a uniform (maximum entropy) distribution over actions is achieved. It is for this reason, that we refer to τ as a post-decision noise parameter: the decision of the agent is given by $\arg\max_{a \in A_{s^2}} v(a, b)$ and deviations from this decision are (by definition) realized after the decision. In other words, the embedding of the agent-specific valence functions in a softmax operation allows for a clear distinction between a deterministic signal (the agent model theory) and observation noise (deviations from the theory modeled by means of the softmax operation). Note that if the latter is interpreted as "random exploration" all deviations from the deterministic prediction of an agent model are interpreted as "exploration", which from an experimenter's perspective is clearly not the case: human participants may adhere to a deterministic decision policy, but on occasion deviate from this policy due to slips of attention, momentary changes in decision policy, working memory lapses and the like. All these random deviations are summarized in the conditionally-independent softmax embedding of the agent value functions, just like the multivariate normal distribution with spherical covariance matrix allows for modeling the sum of arbitrary many random influences from the perspective of the central limit theorem.

How then could random exploration, as suggested by the reviewer, be investigated? Clearly, the statistical modeling behavioral data requires two sources of random variation, one being inherent to the agent in the form of a random decision rule and one modeling observation noise and allowing for deviations from the agent's random decision rule. As such a framework operates based on latent random variables, it requires a filtering approach for model estimation and evaluation (cf. Ostwald et al., 2014b). However, while we share the reviewer's interest in arbitrating between non-random and random exploratory decisions, such an endeavor is beyond the scope of the current study.

In summary, and returning to the reviewer's specific comments,

(1) for the agent models of interest in the current study, there is no need to measure "beliefs about values" as action selection on trial t is predicted deterministically based on observations $o_{1:t-1}$ and actions $a_{1:t-1}$ without intervening latent random variable realizations the state of which could be measured (or, alternatively, integrated out),

(2) the parameter τ is best interpreted as noise parameter which quantifies the degree of deviation between observed participant behavior and the agent decision rule predictions,

(3) the softmax operation is part of the behavioral data analysis, but not the agent models, as random exploration is not investigated here and would required a filtering approach to model estimation and evaluation,

(4) all of the above is documented in the "Data analysis model" and "Model estimation" sections, which we thus deem essential for the documentation of our approach,

(5) the free parameters of the models are clearly specified on lines 460 - 463:

For agents C2, A1, and A2, the conditional log likelihood function is a function of the softmax operation parameter τ only. Thus, for these data analysis models $\theta := \tau$. For agent A3, the conditional log likelihood function is a function of τ and the action valence weighting parameter $\lambda \in [0, 1]$ (...). Thus, for this data analysis model $\theta := (\tau, \lambda)$.

(6) Finally, with respect to the reviewer's comment on conducting a correlation analysis between the noise parameter τ and participant behavior, we are afraid we are unsure as to how such an analysis would further our insights. We note that all agent models are fit to the participants behavior on a trial-by-trial basis and thus their mutual covariation has already been assessed.

The model and parameter recovery could be moved to after the model comparison or to the supplementary information section. You include a model recovery plot for the winning model in Fig. 6c anyways.

We thank the reviewer for this comment, but disagree. As the reviewer has correctly noted, the development and validation of novel computational behavioural models is one of the two main aspects of our work (the other being the development of a novel behavioural paradigm and the presentation of a behavioural data set). In light of this, we believe it is of key importance to first study the model some depth in simulations, communicate what can and cannot be done with a given model, and only then apply to behavioral data to make (modest) inferences about human cognition. We believe that applying a novel, complex, little understood, and unvalidated computational model directly to behavioural data and potentially supporting it by few postdictive checks comes at the cost of reproducibility and building upon our work. The chosen form of presentation (validation in simulated data first, application to behavioural data second) also serves to emphasize that, from our perspective, code validation should always precede code application in the computational behavioural sciences.

S3 could be relegated to the OSF repo, since I doubt anyone would want to transcribe a table by hand these days

We thank the reviewer for this comment, which made us realize that the machine-readable forms of the experimental state sequence had not been uploaded to the OSF repository. They are now available in the repository's Python\Utilities\Ch_opt_seq folder (e.g., `https://osf.io/q2e5s/`). Supplement S3 is the human-readable version of the experimental state sequence and meant for inspection, not transcription.

S4. I wonder if much of the math could be simplified by using conjugate priors (e.g., beta distributions) instead of the more general context described here.

We thank the reviewer for highlighting the details of belief state formation. As stated on p. 11. line 355-356, the aim of Supplement S4 is to prove the validity of eq. (15) of the main text. Equation (15) provides a recursive solution for the sequential evaluation of the agent's observation-conditional an action-dependent state probability distribution $p^{a_{1:t-1}}(s_t^1|o_{1:t-1})$, where s_t^1 takes values $\{1, 2\}$ (cf. the formulation of our task model). As described in the model formulation section, the prior for this distribution on trial t is $p^{a_{1:t-2}}(s_{t-1}^1|o_{1:t-2})$ for trials $t = T, ..., 2$ and $p(s_1^1 = 1) = p(s_1^1 = 2) = 0.5$ for trial $t = 1$. In other words, the sequence of observation-conditional and action-dependent state probability distributions is a sequence of probability mass functions on the outcome space $\{1, 2\}$ of the discrete random variables $s_t^1, t = 1, ..., T$. Beta distributions are distributions of continuous random variables with outcome space $[0, 1]$ and can be used in sequential Bayesian updating schemes to model the (action-independent) observation-conditional reduction of uncertainty about probabilities (e.g. Ostwald et al., 2012, 2015a). We neither consider continuous random variables with outcome space $[0, 1]$ nor modeling uncertainty about the inferred state probabilities here, but consider an action-dependent sequence of distributions. A recursive Beta-Bernoulli scheme does thus not apply for the agent model developed here.

Reviewer #2

The paper introduces a novel explore-exploit choice task in which feedback is withheld for some but not all choices. Using a maximum likelihood modeling approach, the paper finds that participants' choices are best described by a cognitive mechanism that assumes Bayesian belief updating of reward and information gain. The paper addresses an important topic and uses suitable computational modeling techniques. In short, the paper has clear strengths. My main concerns regard conceptual, normative, and methodological aspects of the work.

We thank the reviewer for their constructive feedback on our manuscript.

Conceptually, the relevance of the paradigm to real-world cognition needs to be drawn out more clearly in order to increase the potential impact of this work. The paper presents a somewhat unusual task in which the absence of informativeness is fully randomized across the learning situation. Specifically, on half of the trials, choosing the option with the higher reward rate also yields feedback about the outcome of participants' decisions; on the other half, the option with the lower reward rate provided outcome feedback. Of course, this is an interesting laboratory task for separating exploration and exploitation, but I am hard-pressed to see the ecological validity of a situation in which feedback is lacking in an entirely unsystematic fashion. In real world decisions, information or feedback are often systematically missing for particular (often novel) options. Indeed, the one example the paper provides is a situation in which choosing a novel medical treatment provides information on this treatment, whereas sticking with an existing treatment affords no information about the new drug and somewhat less novel information about the existing treatment. This situation, however, maps more closely onto a typical two-armed bandit situation in which feedback is provided for selected actions only.

We thank the reviewer for this comment and agree that in the original version of the manuscript we only briefly touched upon a real-world scenario that exhibits some structural characteristics of the reinforcement learning problem considered in our study. Furthermore, as the reviewer correctly remarks, the experimental task of the current study has been developed primarily as a novel laboratory task for separating exploration and exploitation in a manner that goes beyond previously established paradigms, such as the classical bandit paradigm and the pure exploration paradigm.

However, we are convinced that humans at least on occasion face the decision dilemma that is modelled by our paradigm, i.e., having to choose between a potentially lucrative and informative or a potentially detrimental and non-informative course of action (as modelled by Trial Type I); or between a potentially lucrative, but non-informative, or a potentially detrimental, but informative, course of action (as modelled by Trial Type II). Consider for example the situation of a patient exhibiting Corona symptoms during early stages of the global pandemic. Under the assumption that there exist highly reliable, medically administered Corona tests and self-administered Corona tests are unavailable (or highly unreliable), the patient faces the following decision dilemma: on the one hand, the patient may stay at home, thereby minimizing the risk for virus transmission (an unobserved lucrative result), but at the cost of obtaining no information about their personal state of infection. In other words, the patient could pursue a potentially lucrative, but uninformative course of action. On the other hand, the patient may use the public transport system to undergo a medically administered Corona test. In this case, the patient incurs the risk of transmitting the virus (an unobserved detrimental result), but obtains reliable information about their personal state of infection. In other words, the patient could also pursue a potentially detrimental, but informative course of action. Depending on the context, the situation may also be inverted, resulting in a Trial Type I-like situation: assuming that the patient has a medically trained relative who can administer a Corona test in the comfort of the patient's home and that the local public health system is of poor quality with largely unqualified newly trained officials administering the official Corona tests, the patient could also be in the situation of choosing between a potentially lucrative and informative as well as a potentially detrimental and uninformative course of action. In summary, we believe, that while the paradigm developed in our study does admittedly not describe the most basic reinforcement learning problem, real-world decision problems that are characterized by trade-offs between obtaining rewards or information and in which information lacks unsystematically for both high and low reward rate options

can be identified. In accordance, we rewrote the corresponding paragraph of the Introduction as follows:

While a large variety of real-life decision-making problems can be modelled with the classical bandit and the pure exploration paradigms, neither variant is suited to model a class of decision-making problems, in which each available action yields positive or negative reward, while only some actions also yield information about the problem's reward structure. Consider for example the situation of a patient who exhibits Corona symptoms during the global pandemic. Assuming that reliable and medically administered Corona tests are available, while self-administered Corona tests are not (as has been the case in many countries in the first year of the pandemic), the patient faces the following decision dilemma: on the one hand, the patient may home-quarantine, thereby minimizing the risk for virus transmission, but at the price of not obtaining information about their own state of infection. Alternatively, the patient may use public transport to undergo a medically administered Corona test, thereby incurring the risk of further transmitting the virus, but obtaining reliable information about their personal infectious state. Situations of this type are similar to the ones modelled with the classical bandit paradigm as each action has a positively or negatively rewarded consequence (which in the example is of societal nature). Importantly, however, in this situation reward-relevant information is detached from reward for one of the actions (home-quarantining), akin to the pure exploration paradigm. Consequently, such a decision situation poses a more pronounced exploration-exploitation dilemma then both classical bandit paradigms and the pure exploration paradigm, because the decision maker is forced to explicitly evaluate the benefit of information gain against the benefit of reward gain. The aim of the current study is to computationally characterize human sequential decision making in such problems.

Moreover, the authors claim that their task provides a very different perspective on the problem of exploration-exploitation trade-offs to the more commonly used bandit problems or the previously used observe-or-bet task but exactly what benefit this perspective provides is not drawn out.

We thank the reviewer for this comment, but note that in fact we did not claim that the task developed for our study provides a very different perspective on the problem of exploration-exploitation trade-offs, but rather that it complements previously used paradigms for studying human explore-exploit behaviour. To address the reviewer's comment, we have refocussed the first section of the Discussion that addresses the relationship between the behavioral task developed for the current study and previously employed paradigms to study human exploration-exploitation behavior.

As the first of our contributions, we introduced an information-selective reversal bandit task suitable to model a class of sequential decision-making problems, in which information about the conferred reward is not available for every action. In the following, we briefly discuss the key aspects of our task with respect to other tasks in the literature. As alluded to in the Introduction, previous research primarily employed either classical bandit paradigms or pure-exploration paradigms to study sequential decision making under uncertainty (Bubeck et al., 2009; Hertwig and Erev, 2009; Ostwald et al., 2015b; Sutton and Barto, 2018; Wulff et al., 2018). We consider these paradigms in turn.

In the classical bandit paradigm, the deciding agent chooses between a set of 'arms' (Berry and Fristedt, 1985). Similar to our task, each arm confers reward according to its associated reward distribution, and, in contrast to our task, each arm confers also information about its expected reward value. A drawback of this design is that the co-occurrence of reward and information evokes a confound between an action's value estimate and the associated uncertainty: as people tend to favor the action with the higher value estimate, for that action the associated uncertainty becomes smaller - simply because for that action more reward observations are made. This makes it difficult to uncover uncertainty-guided exploration in the classical bandit paradigm (Dezza et al., 2017; Gershman, 2018; Wilson et al., 2014). The task employed in the current study circumvents this problem by adopting a symmetrical reward structure of the actions: the probability of the positive reward for the lucrative action is identical to the probability of the negative reward for the detrimental action. Likewise, the probability of the negative reward for the lucrative action is identical to the probability of the positive reward for the detrimental action. In this manner, each reward observation following the informative action confers the same amount of information about the expected reward value of both the lucrative and detrimental action. Furthermore, as in each trial information is randomly coupled with either the lucrative or the detrimental action, our task arguably evokes a more explicit exploration-exploitation dilemma than the classical bandit paradigm, in particular on trials on which participants face decisions between potentially lucrative, but non-informative and detrimental, but informative actions.

The pure-exploration paradigm models sequential decision-making problems in which an action either confers information or reward. In the classical variant of this paradigm, an action that confers reward automatically terminates the task. In an extended variant, referred to as the 'observe-or-bet task' (Blanchard and Gershman, 2018; Navarro et al., 2016; Tversky and Edwards, 1966), the deciding agent can freely switch between an 'observe action' that confers information and 'bet actions' that confer reward. Specifically, 'observe actions' return information about the expected reward of the bet actions, but do not return reward. 'Bet actions', on the other hand, return rewards according to their associated reward distributions, but no information. Similar to the bet actions, in our task one of the available actions confers only reward, but no information. However, in our task, the other available action does not only confer information, as the 'observe action' does, but it also confers reward. Therefore, while exploration and exploitation are temporally separated in the observer-or-bet task, they have to be balanced simultaneously in our task.

In summary, the task proposed in the current study complements the experimental arsenal for studying human exploration-exploitation behavior in the following sense: in contrast to the classical bandit paradigm, the exploration-exploitation dilemma on each trial is *explicit*, i.e., participants need to actively decide whether to gather reward and/or information, and do not obtain information as a by-product of reward-maximizing decisions as in the classical bandit task. Further, in contrast to the pure-exploration

and observe-or-bet task, exploration and exploitation phases occur *simultaneously* and not separated in time.

Finally, a crucial element of the task is the occurrence of dynamic reversals in profitability of the options. The expectation of dynamic changes has been shown to influence behavior considerably in other explore-exploit dilemmas (such as "restless" bandits) but again the conceptual significance of this aspect is not fully drawn out and no comparison to a stationary situation is provided.

We thank the reviewer for raising this important issue and in our revised Discussion address the conceptual significance of dynamic reversal in profitability of the options as follows:

In summary, the task proposed in the current study complements the experimental arsenal for studying human exploration-exploitation behavior in the following sense: in contrast to the classical bandit paradigm, the exploration-exploitation dilemma on each trial is *explicit*, i.e., participants need to actively decide whether to gather reward and/or information, and do not obtain information as a by-product of reward-maximizing decisions as in the classical bandit task. Further, in contrast to the pure-exploration and observe-or-bet task, exploration and exploitation phases occur *simultaneously* and not separated in time. Finally, by introducing a constant probability for the reward rate reversal of the choice options, the task does not only evoke an explicit and simultaneous exploration-exploitation dilemma, but also one that is *ongoing*. This aspect is in line with recent studies using classical bandit or observe-or-bet tasks that also adopted non-stationary reward structures (Bartolo and Averbeck, 2020; Blanchard and Gershman, 2018; Chakroun et al., 2020; Navarro et al., 2016; Hampton et al., 2006; Speekenbrink and Konstantinidis, 2015) and emulates the ever-changing reward structure of real environments.

On the normative side, I would have hoped to see a rational analysis of behavior in the novel task. That is, what choice behavior would be optimal in this task? How should people optimally tradeoff information seeking and exploitation? What learning dynamics should be expected under optimal behavior? And what role do prior expectations about the dynamic reversals play for optimal behavior?

We thank the reviewer for this comment and share their enthusiasm for these interesting questions. However, the derivation of an optimal choice policy by means of theory of partially-observable Markov decision processes is beyond the scope of the current work. By explicitly specifying the task and agent models, as well as sharing the behavioral data, we believe to have laid the foundations for such an undertaking, but given the extend of the current manuscript have to leave its implementation to future work. Together with the remaining comments by the reviewer, we address this point in a new paragraph of the Discussion section (see below).

Methodologically, I felt that the model comparison included heuristic agents that were perhaps too simplistic and inflexible to be competitive. Specifically, the win-stay, lose-shift rule was implemented deterministically such that a switch always occurs after a loss and a repeat action always occurs after a win. This assumption may be too restrictive and an alternative variant of this heuristic implements probabilistic, parameterized switch and stay rates after losses and wins, respectively. Indeed, such probabilistic win-stay, lose-shift strategies have been shown to account well for behavior in related choice tasks (see e.g., Worthy, Otto, & Maddox, 2012). Similarly, the baseline model allows only for strict guessing with probability of .50 instead of allowing for a constant response rate with any probability. Additionally, I was wondering why no reinforcement learning agents were included in the model comparison. Evidence accrual via reinforcement learning represents a central concept in the explore-exploit literature and a comparison to such a more complex form of responding (beyond heuristics) would have been instructive.

We thank the reviewer for these interesting suggestions. Clearly, no behavioural model set is ever complete. For the current work, we focused on developing and implementing a new behavioral task, acquiring behavioral data, and developing and evaluating a straightforward set of agent models that is guided by the idea of contrasting and integrating belief state-based exploitation and exploration. Given the comprehensive documentation of our study and the availability of the ensuing data, we invite the computational modeling community to use these data to develop and test many additional agent models which may or may not capture the observed human behavior better than the models proposed here. Together with the remaining comments by the reviewer, we address this point in a new paragraph of the Discussion section (see below).

Finally, I felt that 24 participants represent a rather small sample size for establishing a novel paradigm and drawing robust conclusions about the cognitive mechanisms underpinning behavior in this task. I

understand that the sample was collected as part of an fMRI study, but these data are not included in the current paper.

We thank the reviewer for this comment, and while we note that the observed superiority for model A3 when compared to the other agent models of our model space is very robust across the group of participants, agree that in general larger data sets are better than smaller data sets. The current study is hence best understood to be of exploratory character (e.g. Jaeger and Halliday, 1998; Dirnagl, 2020), a fact that we make explicit in the following novel paragraph of the Discussion section.

At last, some limitations of our study along with some suggestions for future research must be acknowledged. First, as for any scientific study, inferences can only be made with respect to the model space, which *per se* is incomplete. Thus, although we provide evidence for a strategy resembling a belief state-based agent seeking to maximize the convex combination of its expected Bayesian surprise and its expected reward, it is very well possible that an agent not included in our model space can better account for the behavioral data. Based on our experimental documentation and the open availability of the data, future research may broaden the model space and could also consider, for example, probabilistic variants of the win-stay-lose-switch agents (Worthy et al., 2012) or agents with non-adaptive learning rates (Wiering and Otterlo, 2012; Rescorla and Wagner, 1972), besides the already mentioned random exploration agents in a probabilistic filtering framework (Ostwald et al., 2014a). Second, on a related note, we here do not derive an optimal agent in the sense of partially observable Markov decision process theory (Bertsekas, 2000; Puterman, 2005; Bäuerle and Rieder, 2011). A model space that comprises an agent implementing an optimal choice strategy would afford the analysis of the participants' behavior from a normative perspective and is thus an interesting endeavor for future research. To facilitate such an undertaking, we have provided a detailed documentation of the task and agents developed thus far. Third, given the modest sample size of our study, the behavioral data reported here are best capitalized on in an exploratory fashion. Together with the implemented open research measures (Ritchie, 2020), we believe to have laid the foundations for reproducing our study and build upon our work.

In addition, we note the exploratory character of the study explicitly in the revised study abstract.

In the current exploratory study, we introduce an information-selective symmetric reversal bandit task to model such situations and obtained choice data on this task from 24 participants.

References

Åström, K. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205.

Bartolo, R. and Averbeck, B. B. (2020). Prefrontal cortex predicts state switches during reversal learning. *Neuron*.

Bäuerle, N. and Rieder, U. (2011). *Markov Decision Processes with Applications to Finance*. Universitext. Springer Berlin Heidelberg, Berlin, Heidelberg.

Berry, D. A. and Fristedt, B. (1985). Bandit problems: Sequential allocation of experiments (Monographs on statistics and applied probability). *London: Chapman and Hall*, 5:71–87.

Bertsekas, D. P. (2000). *Dynamic Programming and Optimal Control*. Athena Scientific, second edition.

Blanchard, T. C. and Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, 18(1):117–126.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, pages 23–37. Springer.

Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. (1994). Acting optimally in partially observable stochastic domains. In *AAAI*, volume 94, pages 1023–1028.

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., and Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, 9:e51260.

Dezza, I. C., Angela, J. Y., Cleeremans, A., and Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific reports*, 7(1):1–13.

Dirnagl, U. (2020). Resolving the tension between exploration and confirmation in preclinical biomedical research. In Bespalov, A., Michel, M. C., and Steckler, T., editors, *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*, pages 71–79. Springer International Publishing, Cham.

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

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J. A., Varoquaux, G., and Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3:160044.

Hampton, A. N., Bossaerts, P., and O'doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26(32):8360–8367.

Held, L. (2008). *Methoden Der Statistischen Inferenz*. Spektrum Akademischer Verlag, Heidelberg.

Hertwig, R. and Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, 13(12):517–523.

Jaeger, R. G. and Halliday, T. R. (1998). On Confirmatory versus Exploratory Research. *Herpetologica*, 54:64–66.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Lauritzen, S. (2009). Laplace's Method of Integration (available from http://www.stats.ox.ac.uk/~steffen/teaching/bs2HT9/laplace.pdf).

Littman, M. L. (2009). A tutorial on partially observable Markov decision processes. *Journal of Mathematical Psychology*, 53(3):119–125.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Number 7 in Advanced Quantitative Techniques in the Social Sciences. Sage Publications, Thousand Oaks.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA.

Navarro, D. J., Newell, B. R., and Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive psychology*, 85:43–77.

Ostwald, D., Kirilina, E., Starke, L., and Blankenburg, F. (2014a). A tutorial on variational Bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, 60:1–19.

Ostwald, D., Kirilina, E., Starke, L., and Blankenburg, F. (2014b). A tutorial on variational Bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, 60:1–19.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., and Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*, 62(1):177–188.

Ostwald, D., Starke, L., and Hertwig, R. (2015a). A normative inference approach for optimal sample sizes in decisions from experience. *Frontiers in Psychology*, 6.

Ostwald, D., Starke, L., and Hertwig, R. (2015b). A normative inference approach for optimal sample sizes in decisions from experience. *Frontiers in psychology*, 6:1342.

Puterman, M. (2005). *Markov Decision Processes*. Wiley-Interscience.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2).

Rescorla, R. A. and Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In *Classical Conditioning II*, page 18. New York.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies — Revisited. *NeuroImage*, 84:971–985.

Ritchie, S. (2020). *Science Fictions - Exposing Fraud, Bias, Negligence and Hype in Science*. Penguin.

Sarkka, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Speekenbrink, M. and Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, 7(2):351–367.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*, volume 1. MIT press Cambridge.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Tversky, A. and Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, 71(5):680.

Wiering, M.-H. and Otterlo, Y.-S. (2012). *Reinforcement Learning*.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074.

Worthy, D. A., Otto, A. R., and Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(6):1640–1658.

Wulff, D. U., Mergenthaler-Canseco, M., and Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin*, 144(2):140.