Signatures of heuristic-based directed exploration in two-step sequential decision task behaviour.

Brands, A. M., Mathar, D., Peters, J.

Biological Psychology, Department of Psychology, University of Cologne, Germany

Contact: a.brands@uni-koeln.de

Author contributions:

A.M.B. and J.P. developed the model extensions. D.M. provided analytical tools. A.M.B. analysed the data and wrote the paper. J.P. and D.M. provided comments and revisions.

Acknowledgements:

This work was supported by a grant from Deutsche Forschungsgemeinschaft (DFG) (PE1627/5-1 to J.P.)

Abstract

Processes formalized in classic Reinforcement Learning (RL) theory, such as model-based (MB) control and exploration strategies have proven fertile in cognitive and computational neuroscience, as well as computational psychiatry. Dysregulations in MB control and exploration and their neurocomputational underpinnings play a key role across several psychiatric disorders. Yet, computational accounts mostly study these processes in isolation. The current study extended standard hybrid models of a widely-used sequential RL-task (two-step task; TST) employed to measure MB control. We implemented and compared different computational model extensions for this task to quantify potential exploration mechanisms. In two independent data sets spanning two different variants of the task, an extension of a classical hybrid RL model with a heuristic-based exploration mechanism provided the best fit, and revealed a robust positive effect of directed exploration on choice probabilities in stage one of the task. Posterior predictive checks further showed that the extended model reproduced choice patterns present in both data sets. Results are discussed with respect to implications for computational psychiatry and the search for neurocognitive endophenotypes.

Keywords: computational psychiatry, model-based, exploration, two-step task, neurocomputational endophenotypes

Introduction

"When we remember we are all mad, the mysteries disappear and life stands explained." - *Mark Twain*

Or at least it starts to make a whole lot more sense. The notion that mental health is an integral part to all of our lives and may vary over time on a continuous scale lies at the heart of dimensional psychiatry. Such dimensional approaches address the criticism of clear-cut categorical demarcations between mental disorders. In this view, symptoms exist on a spectrum, with sub-clinical symptom variations (e.g. of depressed mood, compulsive or avoidant behaviours etc.) present in the *healthy population* (Insel et al., 2010; Robbins et al., 2012).

Transdiagnostic research specifically tackles the traditional symptom-based categorisation and thereby partitioning of mental disorders. High rates of comorbidity present another common issue raised with regard to the current conceptualisation and point to inherent flaws (i.e. commonly co-occurring diseases might be better understood as one shared rather than two distinct entities; Insel et al., 2010; Dalgeshi et al., 2020). Oftentimes, transdiagnostic and dimensional approaches go hand in hand as they both aim to improve our understanding of mental disorders, and aim to bring these approaches more in line with syndrome-based understanding known from other medical disorders (e.g., one does not have leg pain, but a broken leg; Insel 2014; Conway & Krueger, 2021).

Research into the basic computational processes that may go awry in the case of mental disorders provides important groundwork for these efforts (Adams et al., 2016). Computational psychiatry has identified several key mechanisms which likely cut across traditional diagnostic lines (Montague et al., 2012; Huys, Maia, & Frank, 2016; Insel et al., 2010; Moutoussis, Eldar, & Dolan, 2017). Such computationally derived *transdiagnostic endophenotypes* might better differentiate between mental health and disease than symptom-based conceptualisations (Robbins et al., 2012; Wise & Dolan, 2020, Yip et al., 2022; Conway & Krueger, 2021). All of these recent changes ultimately aim at furthering our understanding of mental disorders, improving prevention, diagnosis and treatment options. The hope is to someday be able to provide *precision medicine,* more akin on the common understanding and handling of "medical" disorders (Insel, 2014).

Reinforcement Learning (RL) theory (Sutton & Barto, 2018) has been of central importance in these efforts and extensively studied (Montague et al., 2012; Huys et al., 2021; Wise & Dolan, 2020). Here, two processes have emerged as promising computational endophenotypes mapping onto (sub-) clinical variation in symptoms - *model-based* control and *exploration* behaviour (Goschke 2014; Addicot et al., 2017). Regarding model-based control, behaviour is thought to depend on at least two systems: a *model-free* (MF) system, which selects actions based on past reinforcement and a *model-based* (MB) system that uses a model of the environment to select actions based on a computation of their predicted consequences (e.g. Balleine & O'Doherty, 2010; Daw et al., 2011; Daw & O'Doherty, 2014). The exploration-exploitation trade-off (Addicot et al., 2017; Sutton & Barto, 2018) refers to the process of balancing between selecting novel courses of action (exploration) and doing what has worked in the past (exploitation; Daw et al., 2006; Gershman, 2018, 2019). Here, at least two strategies have been discussed (Gershman, 2018; Wilson et al.,2014; 2021): choice randomization (random exploration), e.g. via SoftMax or epsilon-greedy choice rules (Sutton & Barto, 2018) and directed exploration, which involves the specific selection of options that maximize information gain (Wilson et al., 2021). Interestingly, despite the fact that both model-based control and directed exploration have been conceptually linked to the goal-directed system, both processes have largely been studied in isolation.

For more than a decade, the *two-step task* (TST, Daw et al., 2011), as well as various variants thereof, have been the key paradigm in the study MB and MF contributions to behaviour in human studies. This task has also been a central instrument in computational psychiatry (e.g. Voon et al., 2017; Patzelt et al., 2019; Dolan & Dayan, 2013). The TST is a sequential RL task, consisting of two stages, which each involve a choice between two options (Daw et al., 2011, Figure 1). First-stage choices lead to one of two different second stages (S2, Figure 1) with either high probability (common transition, 70%) or low probability (rare transition, 30%). Transition probabilities are reversed for the other S1 option (Figure 1). Second stage options are then associated with either drifting reward probabilities (*classic version*, see e.g. Daw et al., 2011; Gillan et al., 2016) or drifting reward magnitudes (*modified version*, see. Mathar et al., 2022). Following the S2 choice and outcome presentation (c.f. Figure 1) a new trial begins and participants are set back to S1.

Importantly, MB and MF control make different predictions for S1 choices as a function of reward and transition on the previous trial: a purely MF agent would repeat previously rewarded S1 actions, regardless of transition. In contrast, a MB agent "knows" about the transition structure, and would therefore switch to the other S1 option following the experience of a high reward after a rare transition, and repeat the same S1 choice following the experience of a low reward after a rare transition. In this way, the TST is thought to allow for an estimation of the relative contribution of each system (Daw et al., 2011; Otto et al., 2013a).

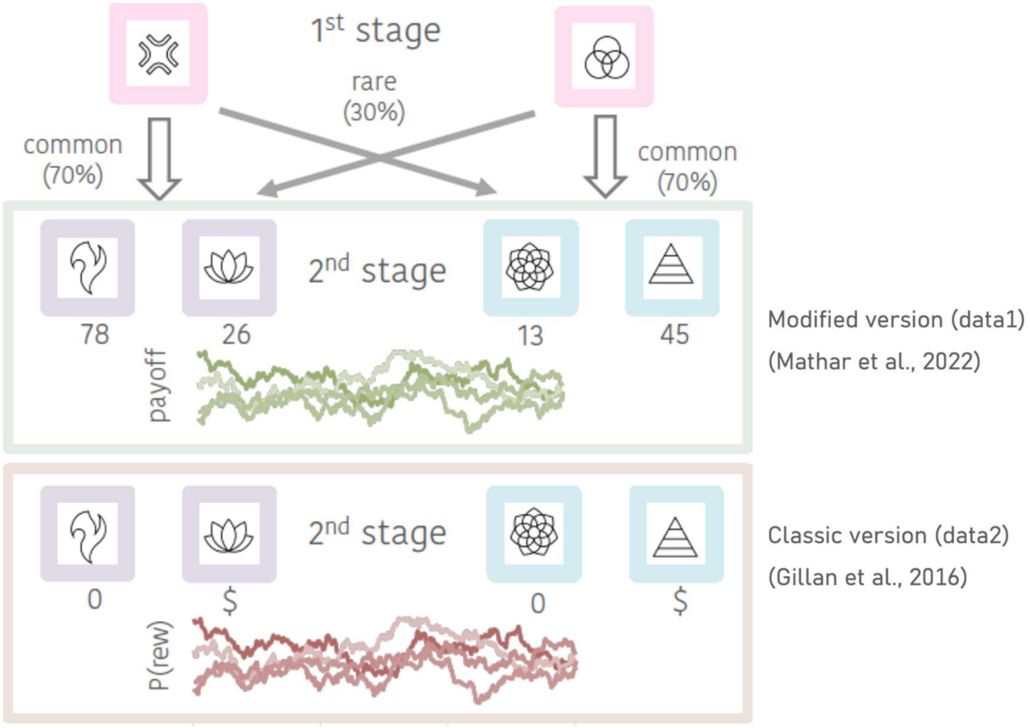


Figure 1. Outline of the two-step task (TST). Transition probabilities from the first stage to the second stage remain the same in both versions of the task. The second stage with a green frame depicts the modified task version employed in data set *data1* (Mathar et al., 2022): after making a S2-choice subjects receive feedback in the form of continuous reward magnitudes (rounded to the next integer).

The lower S2 stage (orange frame) depicts the classic version (used in data set *data2*; Gillan et al., 2016), in which the S2 feedback is presented in a binary fashion (rewarded vs. unrewarded based on fluctuating reward probabilities).

The distinction between MB and MF is conceptually closely related to *goal-directed* vs. *habitual,* behavioural control (Daw et al., 2006; Daw & O'Doherty, 2014; Kool, Cushman, & Gershman, 2018). Despite growing criticisms of this simplified view (see e.g. Akam et al., 2015; Collins & Cockburn, 2020; Feher da Silva & Hare, 2018), this terminology is widely used interchangeably (Miller, Shenhav, & Ludvig, 2019).

A reduction in MB control and resulting over-reliance on the MF system is thought to underlie habitual and/or compulsive behaviors characteristic of several mental disorders (e.g. Voon et al., 2015). Many studies since have indeed provided empirical support for this idea, showing reduced MB control in several patient groups spanning schizophrenia (Culberth et al., 2016), substance use disorders (e.g. Reiter et al., 2016; Sebold et al., 2017), pathological gambling (e.g. Bruder et al., 2021; Wyckmans et al., 2019), eating disorders (Foerde et al., 2021; Reiter et al., 2017), and obsessive-compulsive disorder (OCD; Gillan et al., 2011, Gillan & Robbins, 2014; Brown et al., 2020). Similar effects were observed for sub-clinical variations in symptom severity (Gillan et al., 2016; Seow et al., 2021). Reduced MB control might thus constitute a promising transdiagnostic endophenotype that might more closely relate to real world behaviour than traditional clinical categorizations (Ferrante & Gordon, 2021; Maia & Frank, 2011).

As noted above, directed exploration shares some conceptual features with MB control. Both are assumed to depend on more elaborate computations (Daw & O'Doherty, 2014; Gershman & Daw, 2012; Wilson et al., 2021; for diverging views regarding the dichotomy of MB & MF control see e.g. Akam et al., 2015; Doody et al., 2022; Miller et al., 2019). However, directed exploration might also rely on simpler heuristic-based exploration strategies (Fox et al., 2020). For example, an agent may utilize a simple proxy measure of environmental uncertainty (and therefore of potential information gain) rather than a precise model of environmental dynamics, which is often assumed in e.g. Kalman Filter models (e.g. Daw et al., 2006; Speekenbrink, 2022; Chakroun et al., 2020). Reductions in directed exploration have been observed in a range of mental disorders, including substance use disorders (SUD; Addicot et al, 2013; Morris et al., 2016; Smith et al., 2020), gambling disorder (Wiehler, Chakroun, & Peters., 2021), OCD, Depression (Blanco et al., 2013) as well as anxiety (Aberg et al., 2021; Smith et al., 2020). Similar to model-based control, aberrant directed exploration might constitute a potential transdiagnostic endophenotype in computational psychiatry (Addicot et al., 2017).

Interestingly, the TST shares several key features with tasks traditionally used in the study of directed exploration. For example, similar to restless bandit tasks used to study exploration (Daw et al., 2006; Sutton & Barto, 2018; Speekenbrink, 2022; Chakroun et al., 2020; Wiehler et al., 2021), rewards (or reward probabilities, depending on task version) in S2 of the TST drift according to Gaussian random walks. Potential information gain in S2 of the TST is therefore a function of sampling recency of S2 options. Given the structural similarities between S2 of the TST and restless bandit problems, it thus seems natural to hypothesize that similar directed exploration processes might contribute to TST behaviour. However, there have only been isolated attempts to integrate these concepts (Gijsen et al., 2022). A further issue relates to recent concerns regarding the interpretation of TST results. Feher da Silva and Hare (2018; 2020) argued that MB measures from the TST only reflect one specific model-based strategy (see also Toyama et al., 2017;2019). Participants might well utilize different and/or additional models of the task that constitute a “model-based” strategy, but which are not directly measured using the simple MB vs. MF dichotomy embedded in standard hybrid models (Collins et al., 2017; Collins & Cockburn, 2020; Feher da Silva et al., 2022). To resolve this issue, some researchers have resorted to specifying task features and experimental procedures which allow for a more precise interpretation of results from classical analyses (see e.g. Kool et al., 2016). Another way of tackling this issue is to specify and implement proposed additional processes within computational models, which can subsequently be tested with regard to their fit to empirical data.

The present study follows the latter approach. We had the primary aim to extend standard hybrid models of MB and MF control in the TST by incorporating directed exploration mechanisms. To this end, we implemented and compared several potential candidate mechanisms, including more elaborate mechanisms of uncertainty tracking (Kalman, 1960; Daw et al., 2006) as well as simpler heuristics (Fox et al., 2020). We tested and compared our models in two independent data sets, a variant of the TST with drifting reward magnitudes in S2 (*data1*, Mathar et al., 2022) as well as the classical TST with drifting reward probabilities (and binary payouts) in S2 (*data2*, Gillan et al., 2016). A wealth of empirical data from the TST have already been acquired, many in clinical groups. These data are often available for re-analysis. Therefore, investigations into additional computational mechanisms that might be reflected in these data could proof valuable for the field.

Methods

Participants and Task Versions

We evaluated all models on the basis of a re-analysis of two independent existing data sets. The first data set (*data1*) encompasses data from 39 healthy, male participants (aged 18-35; *M*= 25.17, *SD*= 3.89) who performed 300 trials of the modified TST version (neutral condition from Mathar et al., 2022). The second data set (*data2*) constitutes a subsample (N=100) from a previously published large scale online study using the classical TST (Gillan et al., 2016).

Data1

The first data set (data1) was obtained in a recent study (Mathar et al., 2022) that spanned two testing sessions and included additional tasks, self-report measures and physiological markers of autonomic arousal, which were not analysed for the current study (for more details see Mathar et al., 2022). Prior to performing the TST, participants received instructions regarding the transition probabilities as well as fluctuating reward structure and performed 20 training trials. In addition, they were informed that the maximum response time was two seconds and that they could obtain an additional 4€ in reimbursement, contingent upon task performance.

This TST version used transition probabilities fixed to 70% and 30% for common and rare transitions, respectively and the reward magnitudes for each S2 option followed independent Gaussian Random Walks (fluctuating between 0 and 100, rounded to the next integer, see Fig. 1). S2 states that were marked by different colours to make them more easily distinguishable. For all analyses, trials with response times < 150ms were excluded.

Data2

Gillan and colleagues (2016) used the original variant of the TST (Daw et al., 2011). Here, reward probabilities of all choice options varied independently according to Gaussian Random Walks, and participants received binary reward feedback (rewarded vs. no reward; Fig.1). Detailed descriptions of the whole sample, exclusion criteria, procedure, additional measures, as well as specifics of the TST version employed can be found in the original publication.

We drew a subsample of N=100 (age: M=34; SD=11; 69% female) from the full sample of Experiment 1 from the original publication (N=548, age: M=35; SD=11; 65% female; for further details see Gillan et al., 2016). This subsample is representative of the whole original sample regarding the self-report measures obtained by the authors. To yield this subsample, we randomly sampled from the original full sample from Experiment 1 until self-reported symptom severity did not significantly differ from those of the full sample. The resulting transdiagnostic symptom-score relating to compulsivity and intrusive thoughts (c.f. PCA analyses reported in Gillan et al., 2016 for more detail) was chosen as a criterion due to its significant association with model-based RL in the original publication.

Model-Agnostic Analyses

As a first step we used logistic mixed effects regression models to analyze *stay-probabilities* for first-stage choices, i.e. the probability to repeat the first stage selection of the preceding trial, depending on the transition (common vs. rare) and reward (rewarded vs. unrewarded) experienced. Such regression models are amongst the most common ways of analyzing TST behavior outside a computational modeling framework. Reward and transition type (rewarded/unrewarded and common/rare were coded as 1/-1 respectively) were entered as (fixed effects) predictors of S1 choice repetition (i.e. perseveration). Individual subjects were entered as random effects. Using the *lme4* package (Bates et al., 2015) this resulted in the following model specification in the R syntax:

pstay ~ rew * trans + (rew * trans + 1 | subj)

Due to the presentation of continuous reward magnitudes in the modified version (data1) we defined outcomes to be *rewarded / unrewarded* (1/-1) relative to the mean outcome over the preceding 20 trials (Wagner et al., 2020; Mathar et al., 2022).

As additional model-agnostic indices of MB and MF behavior, we calculated difference scores (MB_{diff} & MF_{diff}) as proposed by Eppinger and colleagues (2013):

$$MF_{diff} = \left[\, P_{common,rewarded} + P_{rare,rewarded} \right] - \left[P_{common,unrewarded} + P_{rare,unrewarded} \right]$$

and $MB_{diff} = \left[P_{common,rewarded} + P_{rare,unrewarded} \right] - \left[P_{common,unrewarded} + P_{rare,rewarded} \right]$.

Computational Models

Our model space consisted of nine models in total. Two standard hybrid RL models without exploration terms but with different learning rules served as baseline models (Q-Learner, Bayesian Learner). We extended both models with exploration terms that incorporated different ways in which S2 uncertainty might impact first stage choice probabilities (see below).

Learning Rules

Q-Learner

The first model is an adaptation of standard hybrid models (e.g. Daw et al.,2011; Otto et al., 2013b). Here, for each first-stage state-action pair (i.e. choice option), separate MF and MB values (Q_{MF}, Q_{MB}) are calculated in parallel.

MF values in both stages are updated using the TD learning algorithm SARSA (Rummery & Niranjan, 1994), such that MF Q-values of a chosen state-action pair at stage i in trial t are updated according to:

(1) $Q_{MF}(s_{i,t},a_{i,t}) = Q_{MF}(s_{i,t},a_{i,t}) + \alpha_i \delta_{i,t}$ with the RPE:

(2) $\delta_{i,t} = r_{i,t} + Q_{MF}(s_{i+1,t},a_{i+1,t}) - Q_{MF}(s_{i,t},a_{i,t})$.

and a constant learning rate α_i(ranging from 0 to 1) for each stage.

S2 prediction errors are incorporated into S1 value estimates via the second-stage learning rate α_2 (constrained between 0 and 1):

(3) $Q_{MF}(s_{1,t},a_{1,t}) = Q_{MF}(s_{1,t},a_{1,t}) + \alpha_2 \delta_{2,t}$.

While several other models utilize an *eligibility trace* parameter (λ) to propagate S2 RPEs, here we chose to reduce the parameter space and model complexity by instead using the S2 learning rate. This formalization was used for all QL models, while all BL models kept the an additional parameter λ (as there is no α_2, for more detail see below; c.f. Table 1).

We included an additional "forgetting process" for MF Q-values (Toyama et al., 2017, Toyama et al., 2019), such that unchosen Q-values decayed towards the mean according to a decay rate α_3(constrained between 0 and 1):

(4) $Q_{MF}(s_{1,t},\bar{a}_{1,t}) = \alpha_3 * Q_{MF}(s_{1,t},\bar{a}_{1,t}) + (1-\alpha_3) * 0.5$ and likewise for S2 MF values: $Q_2(s_{2,t},\bar{a}_{2,t}) = \alpha_3 * Q_2(s_{2,t},\bar{a}_{2,t}) + (1-\alpha_3) * 0.5$.

Recall that transition probabilities in the model were fixed as follows:

(5) $P(s_B| s_A, a_A)$= 0.7, $P(s_C| s_A, a_B) = 0.7$, or in the alternative case as:

$P(s_B| s_A, a_A)$= 0.3, $P(s_C| s_A, a_B) = 0.3$

with

(6) $P(s_B| s_A, a_B) = 1 - P(s_B| s_A, a_A)$ and $P(s_C| s_A, a_A)$= $1 - P(s_C| s_A, a_B)$.

First-stage Q_{MB}values were then computed as the maximal Q_2values weighted by their respective transition probabilities. Thus, using the Bellman equation Q_{MB} values are defined as:

(7) $Q_{MB}(s_A, a_j)$= $P(s_B| s_A, a_j) \max_{a\in\{a_A,a_B\}} Q_2(s_B, a \quad)$

$+P(s_C| s_A, a_j) \max_{a\in\{a_A,a_B\}} Q_2(s_C, a \quad)$

As a trial ends with the second-stage choice, for S2 only MF values are relevant, such that:

(8) $Q_{MB}(s_{2,t}, a)$ = $Q_{MF}(s_{2,t}, a)$=$Q_2(s_{2,t}, a)$.

Accordingly, $Q_2(s_{2,t}, a)$ updates follow the TD process as described previously for first stage Q_{MF} values (Equation 1), while allowing for a separate learning rate α_2 (also constrained between 0 and 1).

Bayesian Learner (BL)

The second learning rule applied is based on the *Bayesian Learner* (BL) as commonly applied in restless bandit tasks (Daw et al., 2006; see also: Chakroun et al., 2020; Wiehler et al., 2021). Here, the constant learning rate α_2 (see Equation 1) is replaced by a trial-specific learning rate (κ_t) based on the Kalman Filter (Kalman, 1960). The idea is that participants track the changes in the underlying reward means of all choice options, as well as the uncertainty associated with these estimates. In this model, learning rates are then uncertainty-dependent.

For both data1 and data2, the true process underlying the Gaussian random walks of S2 rewards was directly incorporated in the model (for applications in the explore-exploit dilemma see e.g. Daw et al., 2006; Chakroun et al., 2020).

For data1, rewards for each option n at trial t ranged between 0 and 1 (multiplied by 100 and rounded to the next integer for presentation as points in the task; c.f. Methods). These were generated following a Gaussian random walk with mean $\mu_{n,t} = 0.5$, and $\mathrm{SD} = 0.04$ (observation variance $\sigma_o^2 = 0.04^2$). Means for each option independently changed on a trial-to-trial basis.

As participants were assumed to represent the reward-walk dynamics in their internal models, the computational model includes fixed parameters $\hat{\gamma}, \hat{\theta}, \hat{\sigma}_o, \hat{\sigma}_d$, which are set to values approximating those of the underlying random walk: the decay parameter γ = 0.9836, decay centre θ= 0.45, and the diffusion variance σ_d^2= 0.025^2 , and diffusion noise $v_t \sim N(0, \sigma_d^2)$.

Participants start with prior beliefs of a normally distributed reward mean $\hat{\mu}_{n,t}^{pre}$ with variance $\hat{\sigma}_{n,t}^{2pre}$ of a chosen second-stage option n on trial t and update these in light of the reward r_t they receive according to:

(9) $\hat{\sigma}_{n,t}^{2post}$= (1-$\kappa_t$)$\hat{\sigma}_{n,t}^{2pre}$ and

(10) $\hat{\mu}_{n,t}^{post}$=$\hat{\mu}_{n,t}^{pre}$+ $\kappa_t \delta_t$ with δ_t= r_t-$\hat{\mu}_{n,t}^{pre}$.

The parameter κ_t is the Kalman Gain, which serves as the uncertainty-dependent learning rate, similar to α_2 in the original (QL) model. In the updating process of a chosen option the Kalman Gain (just like α) scales the prediction error used for updating the mean reward estimate ($\hat{\mu}_{n,t}^{pre}$). The important difference being that κ_t varies on a trial-to-trial basis depending on the observation as well as diffusion variance:

(11) κ_t= $\hat{\sigma}_{n,t}^{2pre}$ / ($\hat{\sigma}_{n,t}^{2pre}$ + $\hat{\sigma}_o^2$).

Thus, high observation uncertainty yields large values of κ_t and thus increased updating. In contrast, low observation uncertainty yields small values of κ_t and thus reduced updating.

Values of all options are updated between trials according to:

$$(12) \qquad \hat{\mu}^{pre}_{n,t+1} = \hat{\gamma}\hat{\mu}^{post}_{n,t} + (1\text{-}\hat{\gamma})\hat{\theta} \text{ and } \hat{\sigma}^{2pre}_{n,t+1} = \hat{\gamma}^{2}\ \hat{\sigma}^{2post}_{n,t} + \hat{\sigma}^{2}_{d}.$$

That is, between-trial updating describes participants tracking of the assumed dynamics underlying the Gaussian random walks. Due to these between-trial updating dynamics, the forgetting process for S2 Q-values (see Equation 4) was omitted from all BL models.

The Kalman Filter (Equation 12) was implemented as the learning rule for second-stage values and then incorporated into the hybrid model, such that $\hat{\mu}^{pre}_{n,t}$ substitutes the Q-values for second-stage options in Equation 13 above.

Choice Rules

The standard SoftMax function (*SM,* McFadden, 1973; Sutton & Barto, 2018) served as the basis for all choice rules. According to this rule, choices' probabilities scale with the value differences between options.

SoftMax

As proposed by Otto and colleagues (2013b), separate coefficients for Q_{MF} and Q_{MB} were used, rather than a single weighting parameter ω (as done e.g. in Daw et al., 2011 and Gillan et al., 2016). Thus, choice probabilities for action a at the first stage were modelled as:

$$(13) \qquad P(a_{i,t} = a|\ s_{1,t}) = \frac{exp\left[\text{ß}_{MB}Q_{MB}(s_{1,t},a)+\text{ß}_{MF}Q_{MF}(s_{1,t},a)+\text{ß}_{C}rep(a)\right]}{\sum_{a\prime} exp\left[\text{ß}_{MB}Q_{MB}(s_{1,t},a\prime)+\text{ß}_{MF}Q_{MF}(s_{1,t},a\prime)+\text{ß}_{C}rep(a\prime)\right]}\ .$$

The parameter ß_C describes the "stickiness" of first stage choices, i.e. first order perseveration. The indicator function $rep(a)$ equals 1 if the first-stage choice of the previous trial is repeated and 0 otherwise.

At the second stage, choices are driven by MF Q-values only ($Q_2(s_{2,t}, a)$, as described above) scaled by the second-stage inverse temperature parameter ß$_2$, such that:

$$(14) \qquad P(a_{i,t} = a|\, s_{2,t}) = \frac{exp[ß_2 Q_2(s_{2,t},a)]}{\sum_{a\prime} exp[ß_2 Q_2(s_{2,t},a\prime)]}\,.$$

The two baseline models M1 and M3 (BL & QL respectively) used this basic version of the SM. These were extended by incorporating terms that account for directed exploration in first stage choices.

Exploration Bonus (eb)

The following sections describe the different implementations of directed exploration for first stage choices in more detail. Generally, we compared various implementations of an exploration bonus (eb) for first-stage values to capture strategic exploration. The general idea is that participants may seek out uncertain S2 states for information-gain and potential long-term reward maximization. Random exploration, in contrast, is assumed to result from sub-optimal, random deviations from a reward-maximizing decision-scheme (Sutton & Barto, 2018; Wilson et al., 2021).

For all variants of the exploration bonus eb, this bonus was included in the standard SoftMax and weighted by an additional free parameter ϕ, resulting in the following first-stage choice probabilities (for models M2 and M4 to M7):

$$(15) \qquad P(a_{i,t} = a|\, s_{1,t}) =$$

$$\frac{exp[ß_{MB} Q_{MB}(s_{1,t},a) + ß_{MF} Q_{MF}(s_{1,t},a) + ß_C rep(a) + \phi eb(s_{1,t},a)]}{\sum_{a\prime} exp[ß_{MB} Q_{MB}(s_{1,t},a\prime) + ß_{MF} Q_{MF}(s_{1,t},a\prime) + ß_C rep(a\prime) + \phi eb(s_{1,t},a\prime)]}.$$

Models M8 and M9 included two measures of eb, therefore another free parameter ρ was included modelling eb_{trial} in addition to eb_{bandit}, resulting in:

$$(16) \qquad P(a_{i,t} = a|\, s_{1,t}) = \frac{exp\left[\text{ß}_{MB}Q_{MB}(s_{1,t},a)+\text{ß}_{MF}Q_{MF}(s_{1,t},a)+\text{ß}_{C}rep(a)+\,\phi eb_{bandit}(s_{1,t},a)+\,\rho eb_{trial}(s_{1,t},a)\right]}{\sum_{a\prime} exp\left[\text{ß}_{MB}Q_{MB}(s_{1,t},a\prime)+\text{ß}_{MF}Q_{MF}(s_{1,t},a\prime)+\text{ß}_{C}rep(a\prime)+\,\phi eb_{bandit}(s_{1,t},a\prime)+\,\rho eb_{trial}(s_{1,t},a\prime)\right]}$$

.

Overall, we compared four different formalizations of uncertainty that participants might draw upon during directed exploration for S1 decisions (see below). These different types of exploration bonus incorporated transition probabilities analogously to the Q_{MB} values (c.f. Eq. 7 & Eq. 18-20). This formalization was based on previous research efforts providing evidence for the demarcation of at least two separate exploration strategies (Sutton & Barto, 2018; Wilson et al., 2014). Directed (vs. random) exploration is assumed to reflect an effortful, goal-directed strategy, aiming at long-term reward accumulation via maximal information gain. In this way directed exploration and MB control show a large conceptual overlap (i.e. deliberate forward-planning under consideration of environmental dynamics, with a long-term perspective on goal-attainment). Consequently, formalizations of directed exploration in model variants presented here were defined analogous to the MB component. In both components transition probabilities are utilized to weigh the reward and uncertainty estimates (MB and exploration, respectively) associated with S2 options to reflect these assumed deliberate and foresighted aspects.

Uncertainty estimates based on the Kalman filter.

This variant was only included for the BL model (M2), as the Q-learning models lack an explicit representation of uncertainty. Here, the eb was based on the estimated maximum standard deviation of second-stage options (Daw et al., 2006; Chakroun et al., 2020) such that:

$$(17) \qquad eb_{sigma}(s_A, a_j) = P(s_B|\, s_A, a_j) \max_{a\in\{a_A,a_B\}} \hat{\sigma}_{n,t}^{pre}(s_B, a)$$

$$+P(s_C|\, s_A, a_j) \max_{a\in\{a_A,a_B\}} \hat{\sigma}_{n,t}^{pre}(s_C, a).$$

The resulting first-stage choice probabilities were modelled according to Equation 15.

All remaining variants of the exploration bonus were based on simpler counter-based heuristics.

Uncertainty estimates based on a bandit-counter heuristic.

For the first of these ($b_{n,t}(s_B, a)$, *bandit-heuristic)*, participants were assumed to estimate how many of the alternative S2 options they have sampled since last choosing a given option n. Following selection of an S2 option n, the respective counter $b_{n,t}$ is reset to 0. Thus, $b_{n,t}$ of a given S2 option n ranges from 0 (this option was chosen on the last trial) to 3 (all other S2 option were sampled since last sampling this option). Diverging from the formalization in the *BL + sigma* model (M2), here we assumed participants to sum up these counters over both options of the respective S2 stage associated most likely with either first-stage choice (instead of tracking the maximum). The sum of $b_{n,t}$ across associated S2 options were again weighted by their transition probabilities (see M2; Equation 17) resulting in:

$$(18) \qquad eb_{bandit}(s_A, a_j) = P(s_B | s_A, a_j) \sum_a b_{n,t}(s_B, a \quad) + P(s_C | s_A, a_j) \sum_a b_{n,t}(s_C, a \quad).$$

This yielded the models M4 (*BL+ bandit*) and M5 (*QL + bandit;* see Equation 15 for corresponding S1 choice probabilities). Often times uncertainty sum-scores are associated with less goal-directed exploration strategies (i.e. random exploration based on total uncertainty), such as Thompson Sampling. In such cases however, total uncertainty scores (sums) are directly linked to choice stochasticity (c.f. Gershman, 2018, 2019; Fox et al., 2020). In contrast, here the summed uncertainty proxy is incorporated in a more complex model of a decision sequence (exploration boni are sensitive to the transition type and are ultimately assigned to S1 state-action pairs). In this way higher sum scores of given counters associated with a particular S1 action are incorporated in a simplified, yet still model-based way. Moran and colleagues (2019) have furthermore used similar formalizations to describe interactive dynamics and partial overlap of the proposed MB and MF system. The authors provide evidence for the incorporation of rather parsimonious MF-like value estimates (via sum scores) to retrospectively assign credit to previous actions using an internal model of the environment.

Uncertainty estimates based on a trial-counter heuristic.

For models *BL+ Trial* and *QL + Trial* (M6 and M7, respectively) we followed the same logic, with the only difference being that participants were assumed to utilize a trial counter ($t_{n,t}$) as a proxy for uncertainty. This counter heuristic was simply defined as the number of trials since that particular second-stage option was last sampled. The resulting exploration bonus was thus defined as follows:

$$(19) \qquad eb_{trial}(s_A, a_j) = P(s_B | s_A, a_j) \sum_a t_{n,t}(s_B, a) + P(s_C | s_A, a_j) \sum_a t_{n,t}(s_C, a).$$

Analogous to models M4 and M5, action probabilities for first-stage choices were modelled according to Equation 15.

Combined models

Finally, combined models with terms for both trial- and bandit-heuristic exploration were examined, i.e. *BL+ Trial + Bandit* and *QL + Trial + Bandit* (M8 and M9, respectively). The implementation of both exploration mechanisms resulted in first-stage action probabilities according to Equation 16.

Hierarchical Bayesian Modelling Scheme

Table 1 provides an overview of all free and fixed parameters for the models described above. Using a hierarchical Bayesian modelling scheme, subject parameters were drawn from shared group-level Gaussian distributions. This resulted in two additional free parameters $M^x$and Λ^x for each subject-level parameter x. Group-level parameter means (M^x) were assumed to be normally distributed (*M*= 0, *SD*= 10) and standard deviations (Λ^x) were set to follow a uniform distribution (with limits 0 and 10 for λ and all α_i , and an upper limit of 20 for remaining group-level SD parameters). All learning rates ($\alpha_1, \alpha_2, \alpha_3$) as well as the eligibility trace parameter (λ) were then back-transformed to the interval (0,

1) using STANs built in cumulative density function. This was done directly within the model, so that raw subject-level parameter values ranging from -10 to 10 were mapped onto the interval (0,1).

For models M1, M2, M4, M6 and M8, participant's estimates of the random walk parameters ($\hat{\gamma}$, $\hat{\theta}, \hat{\sigma}_o, \hat{\sigma}_d$) were fixed to values reasonably approximating the underlying dynamics. The estimated mean and standard deviation for the reward distribution were initialised at the first trial with $\hat{\mu}_{n,1}^{pre} = 0.5$, and $\hat{\sigma}_{n,1}^{2pre}$ = 0.04^2.

Table 1

Free and fixed parameters for all models.

Model	free parameters	fixed parameters
M1 (BL)	$\alpha, \alpha_3, \lambda$, ß$_{MB}$,ß$_{MF}$,ß$_c$,ß$_2$	$\hat{\gamma}, \hat{\theta}, \hat{\sigma}_o, \hat{\sigma}_d, \hat{\sigma}_{n,1}^{2pre}$, $\hat{\mu}_{n,1}^{pre}$
M2, M4, M6 (BL + EB)	$\alpha, \alpha_3, \lambda$, ß$_{MB}$,ß$_{MF}$,ß$_c$,ß$_2$, ϕ	$\hat{\gamma}, \hat{\theta}, \hat{\sigma}_o, \hat{\sigma}_d, \hat{\sigma}_{n,1}^{2pre}$, $\hat{\mu}_{n,1}^{pre}$
M8 (BL + EB$_{\text{bandit}}$+EB$_{\text{trial}}$)	$\alpha, \alpha_3, \lambda$, ß$_{MB}$,ß$_{MF}$,ß$_c$,ß$_2$, ϕ, ρ	$\hat{\gamma}, \hat{\theta}, \hat{\sigma}_o, \hat{\sigma}_d, \hat{\sigma}_{n,1}^{2pre}$, $\hat{\mu}_{n,1}^{pre}$
M3 (QL)	$\alpha, \alpha_2, \alpha_3$, ß$_{MB}$,ß$_{MF}$,ß$_c$,ß$_2$	
M5, M7 (QL + EB)	$\alpha, \alpha_2, \alpha_3$, ß$_{MB}$,ß$_{MF}$,ß$_c$,ß$_2$, ϕ	
M9 (QL + EB$_{\text{bandit}}$+EB$_{\text{trial}}$)	$\alpha, \alpha_2, \alpha_3$, ß$_{MB}$,ß$_{MF}$,ß$_c$,ß$_2$, ϕ, ρ	

Note. QL refers to the basic hybrid model. BL denotes the altered hybrid model versions with Kalman Filter updates. E = added first-stage exploration bonus; ϕ: parameter that scales the exploration bonus. Note that this parameter remains the same for both exploration bonus variants, regardless of the specific formalisation of uncertainty estimates in a given model. For M8 and M9 an additional free parameter ρ was included to account for additive exploration effects of both heuristics.

All models were implemented using the *STAN* modelling language (version 2.21.0; Stan Development Team, 2019) running in the statistical program R, which was also used for all further analyses (version

3.6.1, R Core Team, 2019). The sampling for each model was done using a Markov-Chain-Monte-Carlo (MCMC) algorithm (no-U-turn sampler NUTS), with four chains running 10000 iterations each, 8000 of which were discarded as warm-up. MCMC methods are based on the generation of a random number sequence (chain) that is used to sample a probability distribution. Parameters estimates with higher (posterior) probability are sampled more often, resulting in a posterior probability distribution. The desired state is reached when a chain has reached equilibrium. $\hat{R}$ is a measure of convergence across chains, indicating the ratio of between-chain to within-chain variance. Here, values of $\hat{R} \leq 1.1$ were considered acceptable.

In a first step, all models were compared regarding their predictive accuracy. As this method only provides a relative comparison between models, posterior predictive checks were performed to gain a deeper understanding. These allow a more detailed insight with regard to predictions made by the model and their ability to accurately portrait the data as well as an indication of possible model misspecifications (Wilson & Collins, 2019).

Open Code

STAN model code will be shared via the *Open Science Framework* upon publication. By making the model code freely available, we wish to facilitate further application and development of this model and further adaptations thereof. Transparently reporting on model specifications also holds the potential of direct comparisons of parameter estimates (e.g. as reported above for the data sets compared here).

Open Data

Both data sets analyzed here are freely available (for links see the respective publications).

Model Comparison

Model comparison was performed using the *loo* package (Vehtari et al., 2023) which provides a measure of predictive accuracy via leave-one-out cross-validation (LOO; Vehtari, Gabry, & Gelman, 2017). To this end the estimated log pointwise predictive density (-elpd) is applied as the criterion of interest. The model with the lowest -elpd score was selected as the best fitting model. While lower values indicate a superior fit, in direct comparisons between models values close to 0 indicate superior fit. As can be seen in Table 3, in these cases the difference in elpd compared to the winning model (*elpd* *diff*) is provided (thus, values close to zero show “smallest” distance to the winning model). In cases in which a more parsimonious model showed overlap with the best-fitting model in terms of the SE of the elpd difference, the more parsimonious model was chosen.

RESULTS

Model-Agnostic analyses

In a first step, we quantified MF and MB contributions using common model-agnostic procedures (see e.g. Daw et al., 2011; Otto et al., 2013a; Gillan et al., 2016) as outlined in the methods section. A linear mixed model of S1 stay-switch behavior using the factors reward and transition type as well as their interaction confirmed the standard effects (Daw et al., 2011; Otto et al., 2013a): in both data sets (see Table 2) we observed a main effect of reward (reflecting MF control) and a reward x transition interaction (reflecting MB control).

Table 2. Results from regression analyses of S1 choice repetition probability.

		Estimate	95% CI	z-Value	p-Value
data1	Intercept	1.35	[1.12; 1.58]	11.48	<.01
	Reward	0.11	[0.05; 0.18]	3.47	<.01
	Transition	-0.07	[-0.14; -0.01]	-2.14	.03
	Reward*Transition	0.47	[0.36; 0.59]	8.10	<.01

data2	Intercept	1.73	[1.53; 1.94]	16.68	<.01
	Reward	0.64	[0.51; 0.77]	9.89	<.01
	Transition	0.02	[-0.03; 0.08]	0.81	.42
	Reward*Transition	0.16	[0.07; 0.24]	3.76	<.01

Note. Reward: main effect of reward type (unrewarded vs. rewarded), commonly interpreted as an indicator for MF control; Transition: main effect of transition type (rare vs. common); Reward*Transition: interaction of Reward and Transition type, commonly interpreted as an indicator for MB control.

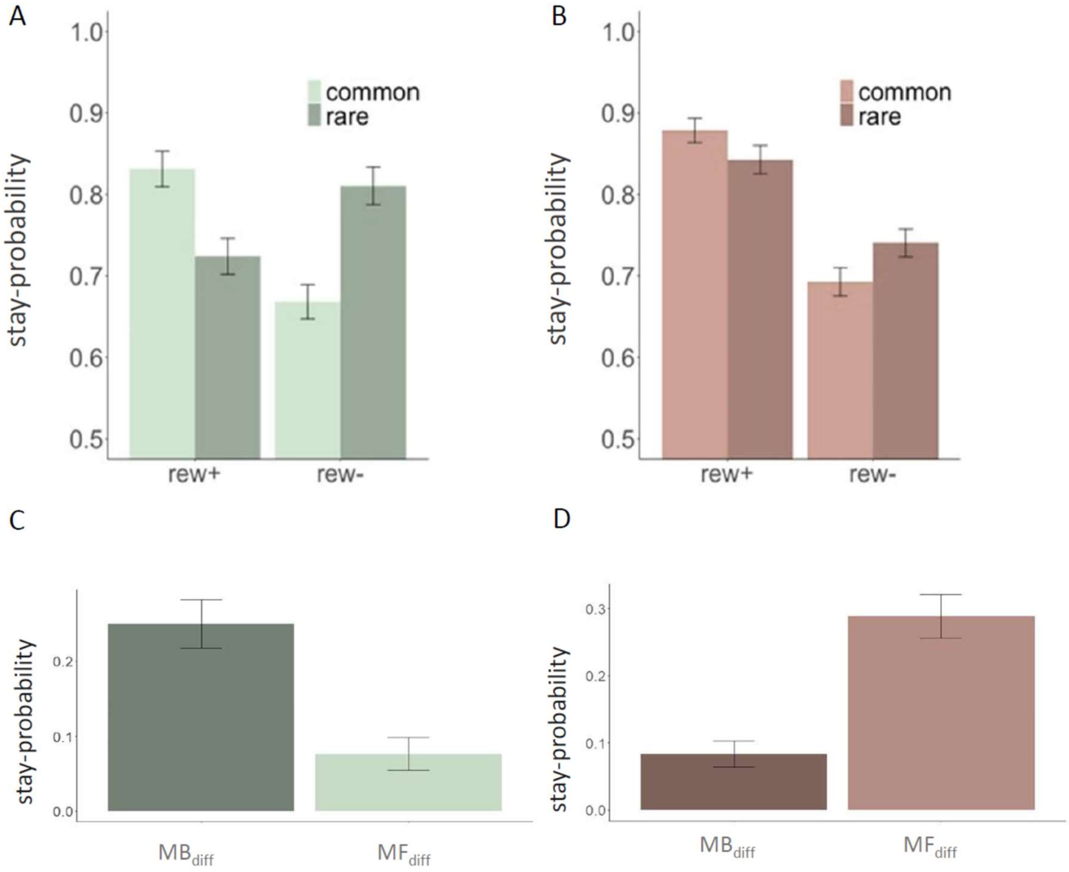


Figure 2. Stay-Probabilities of S1 choices and difference scores. Upper panel: Probabilities for S1 choice repetition as a function of reward (rew+ : rewarded; rew- : unrewarded) and transition type (common/rare) of the preceding trial. Lower panel: MB and MF difference scores as defined by Eppinger et al. (2013), bar heights depict mean scores over all participants, error bars show the standard error. The left plots (green, A & C) shows results from data1; the right plot (orange) shows results from data2.

However, regression results along with visual inspection (Table 2, Figure 2) also suggest differences between task versions (i.e. between data1 and data2). The MF effect was somewhat more pronounced in the data set from Gillan and colleagues (2016; data2), while data1 showed a more pronounced MB effect. This contrast between data1 and data2 was also clearly evident in the respective model-agnostic difference scores for the two effects (Figure 2, lower panel).

Model comparison

In both data sets all parameters (group- as well as subject-level) could be estimated well, as evidenced by the aforementioned convergence measure $\hat{R}$ (all $\hat{R}$ <1.1). Model comparison then was based on the estimated log pointwise predictive density (-elpd). Here, lower absolute values reflect a superior fit. The difference in -elpd (-*elpd* *diff*; c.f. Table 3) is provided in reference to the winning model (which itself thus always has a -*elpd* *diff* of zero). The Q-Learner models consistently outperformed BL models across both data sets (cf. Figure S1).

Within the group of QL models the bandit model (M5) provided the best fit for data2 and was indistinguishable from the combined model (M9) for data1 (i.e., the standard error of the elpd-difference of M5 from the best model M9 included zero, see Table 3, Figure 3). Thus, all further analyses focused on the more parsimonious model (M5, *QL+Bandit*).

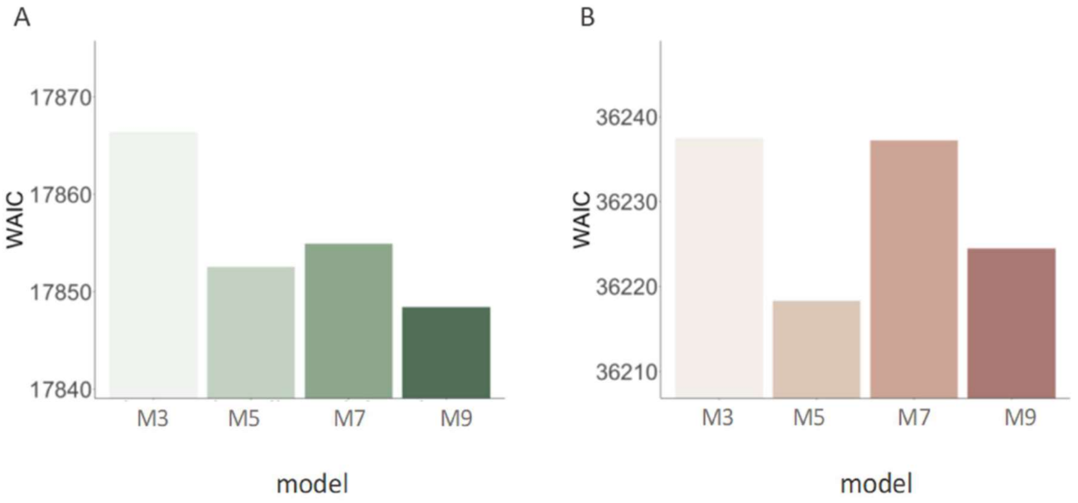


Figure 3. Model Comparison Results via the Widely Applied Information Criterion (WAIC) for all QL Models (M3, M5, M7, M9). Panel A/B green/orange bar plots refer to data1 and data2, respectively.

M3: *QL base* model without added exploration bonus. M5/7: *Base + Bandit/ Trial* refer to the model variants with added heuristic-based exploration bonus using stimulus identity/recency, respectively. M9: *Base + Bandit+ Trial*: model variant with additive combination of both heuristic-based exploration implementations.

Table 3

Results from model comparison of QL-models using Leave-one-out cross-validation (LOO).

Data Set	Model	*-elpd diff*	*se diff*
data1	M3 (Base)	-7.9	6.3
	M5 (Base + bandit)	-0.4	5.0
	M7 (Base + trial)	-5.4	4.5
	M9 (Base + bandit + trial)	0.0	0.0
data2	M3 (Base)	-10.8	8.1
	M5 (Base + bandit)	0.0	0.0
	M7 (Base + trial)	-8.5	7.3
	M9 (Base + bandit + trial)	-4.2	6.9

Note. The difference in the expected log pointwise predictive density (*elpd* $_{diff}$) and standard error of the difference (*se* $_{diff}$). These values show the results of a model comparison using LOO estimates. Each model is compared to the preferred model (M5/M9), as there is no difference between the winning model and itself, values in the first column are always zero.

Posterior Predictive Checks

While results from model comparisons can provide relative support for one computational account over another, posterior predictive checks are required to ensure that a model can reproduce core patterns in the data. We thus simulated 10000 full trial sequences, 8000 of which were discarded as warm-up, yielding 8000 (2000 per chain) simulated data sets per subject. Simulations were performed during fitting the model to the empirical data (i.e. on the basis of possible parameter values sampled during this procedure). As can be taken from Table 4, simulated S1 choices largely reproduced the patterns observed in human data. We repeated the model-agnostic analyses of stay-/switch behavior (shown above for empirical data, c.f. Table 2 Figure 2 and 4) for the simulated data sets. Visual inspection (Figure 4) revealed an underestimation of *"stay"*-probabilities in the simulations when compared to the empirical data. Nonetheless, the overall pattern of stay-switch tendencies as a function of reward and transition were largely reproduced, whereas this was less pronounced for the main effect of reward (MF contribution; Table 2, Figure 2).

Table 4. Proportion of correct S1 choice predictions by the winning model M5.

Data Set	Min	25th percentile	Median	Mean	75th percentile	Max
data1	.510	.631	.758	.742	.840	.9125
data2	.505	.667	.753	.749	.828	.9754

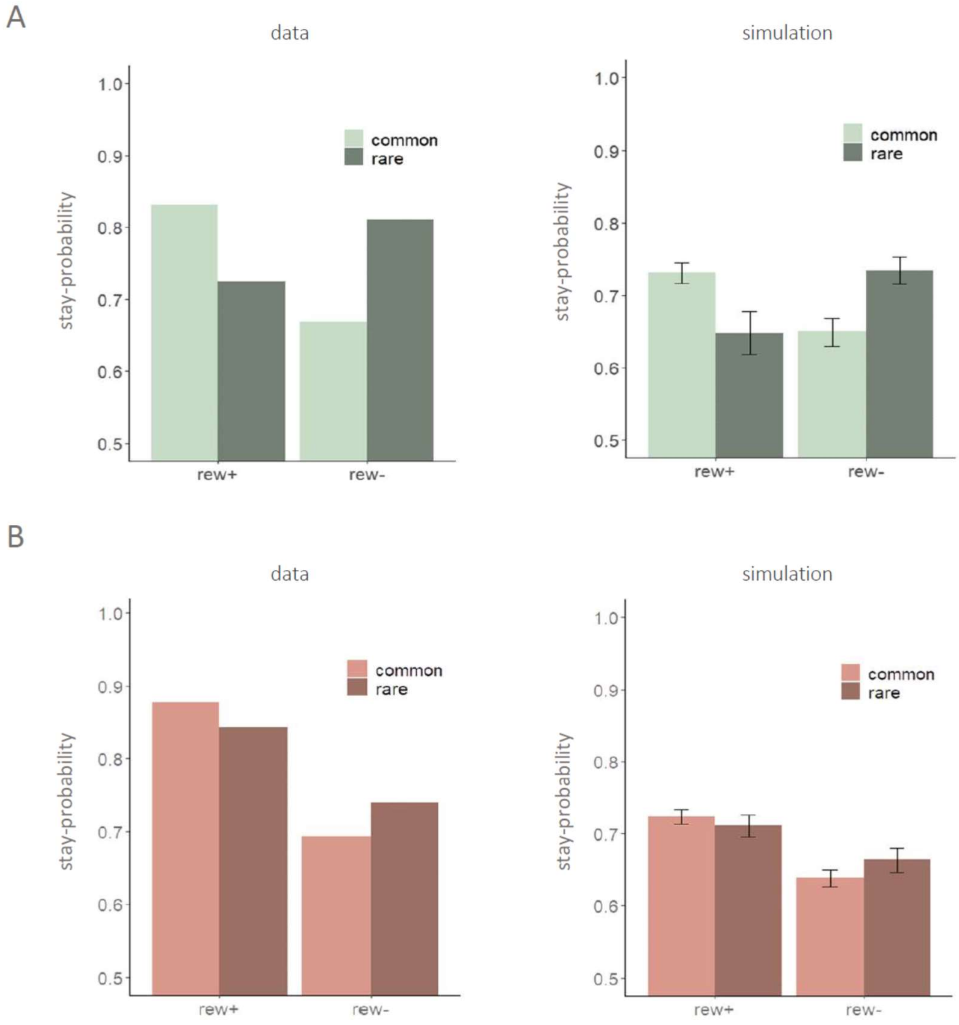


Figure 4. Probabilities of S1 choice repetition as a function of reward and transition type. Y-axis: Stay probabilities for 1st stage choices; data: empirical stay probabilities from data sets data1 (panel A; green) and data2 (panel B; orange). simulation: stay-probabilities from N=8000 simulated choice sequences per subject, derived from the winning model (M5).; rew+/-: previous trial was rewarded (+) or unrewarded (-).; common/rare: previous trial followed a common/rare transition respectively. Error bars in the simulation plots depict the 95% HDI over 8000 simulated data sets.

Posterior Distributions

Figure 5 shows the posterior distributions of group-level parameters underlying S1 choices in the best-fitting model. The group-level mean of the exploration bonus parameter (ϕ) was positive in both data sets and the 95% highest-density interval (HDI) did not overlap with 0 (Table 5, Figure 5). We thus confirmed the predicted positive effect of directed exploration on S1 choice probabilities in both data sets. Furthermore, in both data sets, there was evidence for a substantial perseveration effect (Figure 5).

Posterior point-estimates of all hyperparameter means from the best-fitting model are shown in Table 5. Model parameters confirmed the model-agnostic analyses portraited above: the MB component (ß$_{mb}$) was more pronounced in data1 vs. data2.

Table 5. Posterior Point-Estimates of all Group-Level Parameters of model M5 for data1 and data2.

parameter	Mx		Λx	
	data1	data2	data1	data2
α_1	-0.63	0.32	1.40	0.59
α_2	0.47	-0.11	1.05	0.78
α_3	2.11	0.75	0.64	1.26
ß$_{mb}$	12.26	2.73	10.80	5.26
ß$_{mf}$	1.73	3.00	2.03	2.22
ß$_{persev}$	0.86	1.58	0.52	1.08
ß$_2$	10.13	6.78	4.75	2.91
ϕ	0.15	0.09	0.13	0.08

Note. Posterior point-estimates of group-level means (Mx) and standard deviations (Λx) for data1 and data2 from the winning model (M5; QL +Bandit) for all subject-level parameters x listed in the first column.

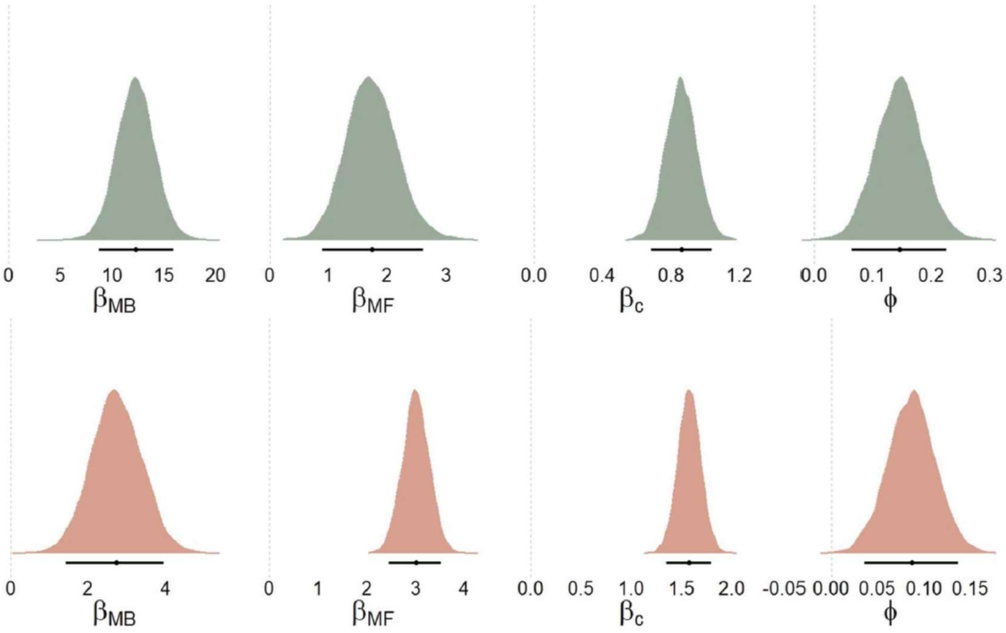


Figure 5. Posterior distributions of Group-Level Means of S1 Choice Parameters. Solid black lines show the 95% highest density interval (HDI) and the black dot depicts the point-estimate of the mean. Panels A and B (green and orange plots) show results on the basis of data sets data1 and data2, respectively.

Recall that in the best-fitting model M5, the uncertainty heuristic (bandit heuristic) ranges between values of 0 to 3 per S2 option, resulting in a range of 1 to 5 for the combined predictor over both S2 options (see Equation 18 above). Recall further that rewards were re-coded to range between 0 and 1 for both data sets. Table 6 illustrates the resulting exploration bonus values (V_{EB}) for four different example trials and parameter values of ϕ = 0.05; 0.1 and 0.15. Imagine a subject choosing option A in S1 (*S1A* first column), followed by a common transition, thus, leading to the S2-state A (*S2A* third column). Since the last visit of state S2A our subject may have chosen two and three alternative S2 options each, so that the *bandit-heuristic counters* of the currently available S2 options have values two and three respectively. Following our model, the sum of both counters (five in this example) is scaled by the transition probability (0.7 as it was a common transition from S1A) and the *alternative case* is added (see Equation 18). In this alternative case (following a rare transition, with probability 0.3 leading to S2B instead) *bandit counters* for the available choice options may have values zero and one respectively (e.g. if these were chosen in the two preceding trials). These resulting model-based uncertainty estimates are then weighted by the individual's exploration parameter ϕ (here shown for exemplary values 0.05, 0.1, and 0.15).

Table 6. Exemplary exploration bonus values for different S1-S2 combinations given plausible exploration bonus parameter values for ϕ

S1-choice	transition	S2 state (*bandit* heuristic counter value)	ϕ	Resulting exploration bonus value (V_{EB})
S1A	Common	S2A (2; 3)	0.05	0.19
			0.1	0.38
			0.15	0.57
S1B	Common	S2B (0; 1)	0.05	0.11
			0.1	0.22
			0.15	0.33
S1A	Rare	S2B (2; 3)	0.05	0.115
			0.1	0.23
			0.15	0.345
S1B	Rare	S2A (0; 1)	0.05	0.17
			0.1	0.34
			0.15	0.6

Note. The first three columns show possible combinations of first state choices and second stage states following either transition type. ϕ: plausible exemplary subject-level parameter values. The last column shows the exploration bonus value for the these S1- S2 sequences for a given ϕ-estimate according to Equations 15 and 18.

A rough insight into the relationship of V_{EB} compared to the weighted MB Q-value (Q_{MB}) of S1 options is shown in Table 7. In data1 the median proportion of V_{EB} in relation to Q_{MB} (V_{EB} / Q_{MB} over both S1 options) was 5%, whereas it was 17% in data2. These example calculations illustrate that, while posterior estimates of the exploration parameter may appear numerically small, they do in fact make meaningful contributions to subjective valuations.

Table 7. Proportion of V_{EB} compared to Q_{MB} across both S1 options

Data Set	25th percentile	Median	75th percentile
data1	.03	.05	.11
data2	.08	.17	.42

Note. V_{EB} and Q_{MB} were calculated on the basis of subject-level parameter estimates derived from M5 (QL + Bandit). Values shown depict V_{EB} /Q_{MB} (c.f. Eq. 18 and Eq. 7 respectively) over both S1 options, which were then averaged across all participants for data1 and data2 each.

Correspondence with model-agnostic analyses

In order to investigate how parameters derived from the model relate to model-agnostic indices of MB and MF behavior as well as to the overall performance, correlation analyses were performed (Figure 6). For this purpose, we applied the same regression model as described for the group analyses to each participant's individual data set, omitting the random effects term, resulting in:

pstay ~ rew * trans + (rew * trans + 1).

Model-agnostic indices of MB ($ß_{rew:trans}$ and MB_{diff}) and MF ($ß_{rew}$ and MF_{diff}) exhibited moderate associations in both data sets (Figure 6). Both effects were likewise associated with the corresponding model-derived parameters ($ß_{MB}, ß_{MF}$). The exploration parameter did not show significant associations with mean rewards earned on this task. In contrast, the MB component showed a moderate to strong association to the mean overall payout in data1 but not data2, confirming that the modified version successfully addressed previous concerns (see Kool et al., 2016). In data2, there was no evidence for this association (see Figure 6, lower panel).

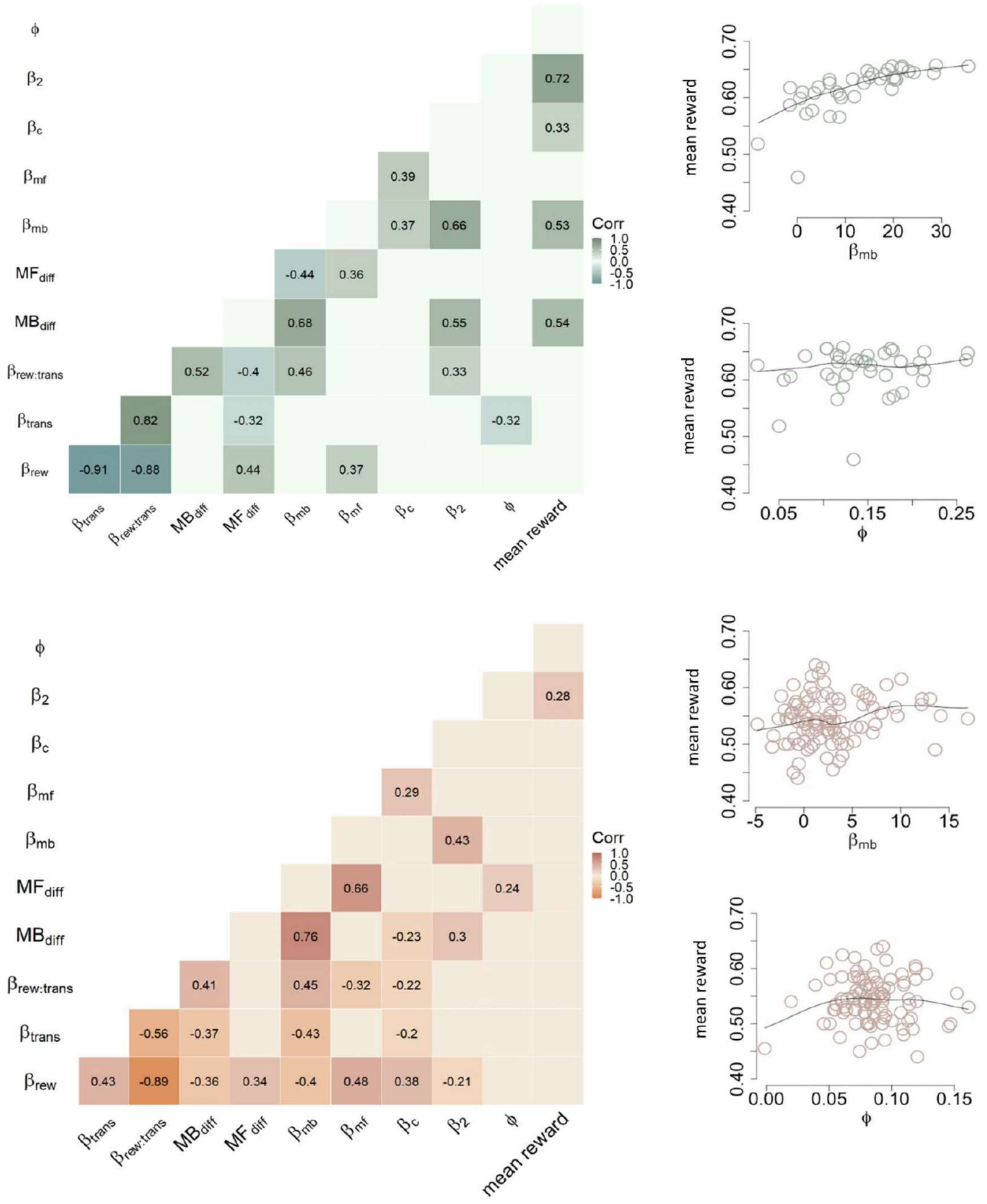


Figure 6. Associations of model-agnostic and model-derived TST indices of MB and MF signatures. Results from correlation analyses of model-agnostic indices of MB and MF influences. Empty tiles (left panel) indicate non-significant associations. ; ß$_{rew}$,ß$_{trans}$,ß$_{rew:trans}$: regression weights for main effects of reward, transition type and their interaction; MB_{diff}, MF_{diff}: differences scores of MB and MF influences on S1 stay probabilities respectively; ß$_{MB}$,ß$_{MF}$: MB and MF S1 choice parameters from the winning model; ß$_{c}$: S1 perseveration parameter; ß$_{2}$: S2 inverse temperature parameter; ϕ: exploration parameter; mean reward: mean reward gained throughout TST (data1: 300 trials, data2: 200 trials). Right panel: association of model-derived MB (ß$_{MB}$) and exploration parameter (ϕ) with mean reward. Circles depict individual participants. All green plots (upper panel) are based on data1, orange plots (lower panel) are based on data2.

Further Validation of the best-fitting model

In a final step, we verified that results for data2 were not due to specific characteristics inherent to the subsample we drew from the data set from Gillan et al. (2016). To this end, we repeated the model estimation procedure for M5 (c.f. Methods described above) using the full sample of experiment 1 (N=548) from the original publication by Gillan and colleagues (2016). Modelling results were comparable to those reported above for the initial subsample (data2). The model converged equally well (all $\hat{R}$<1.1) and posterior estimates of group-level parameters mirrored results reported above (Figure7). Results again confirmed a highly robust effect of S2 uncertainty on S1 choice probabilities (see Figure 7), such that the posterior distribution of the exploration bonus parameter ϕ was robustly > 0.

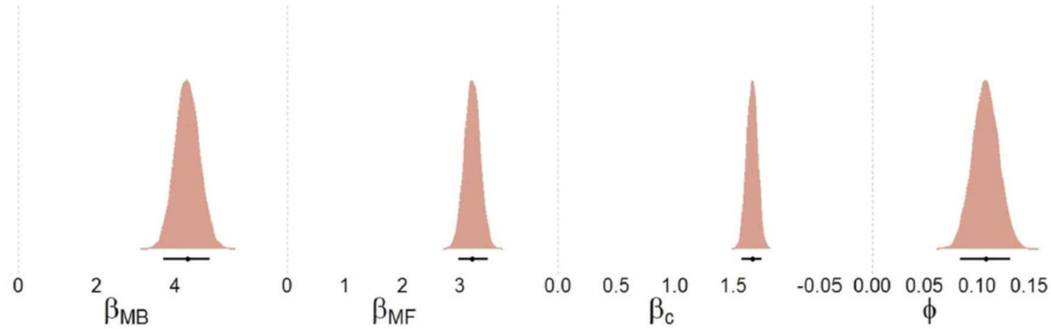


Figure 7. Posterior density estimates of 1st-stage group-level mean choice parameters from the winning model. The lower panel of Figure 5 shows corresponding results based on data2 (N=100), plots shown here are based on the full sample of experiment 1 in the original publication (N=548; Gillan et al., 2016). Black dots indicate the mean point-estimate, black bars cover the 95%-HDI.

DISCUSSION

This study set out to extend existing hybrid models of TST behaviour with mechanisms implementing directed (uncertainty-based) exploration. To this end, we considered varying ways in which uncertainty may guide participants choices in stage 1 of the task, building upon insights from the explore-exploit dilemma (4-armed restless bandit; Daw et al., 2006; Chakroun et al.,2020; Wiehler et al., 2021). In the empirical data analysed here, an uncertainty-dependent learning model (BL-models using a Kalman Filter) did not provide an advantage over a classic Q-Learner algorithm. The addition of heuristic-based uncertainty measures in the choice component for stage 1 of the task, however, significantly improved model fit (see model comparison results). Results were consistent across two independent data sets using two versions of the TST assessed in different settings (laboratory vs. online sample), and confirmed a positive directed exploration effect in TST behavioural data.

We complemented model-based results with more traditional model-agnostic analyses to gain a deeper understanding of how these relate to each other. This additionally aided the interpretation of the best-fitting model, parameter estimates derived from it as well as differences in the results obtained from both data sets.

Uncertainty-based exploration (but not learning) on the TST

While human choice behaviour in the bandit task is typically better described by models that incorporate a Kalman-Filter as a learning component (vs. a constant learning rate; see e.g. Daw et al., 2006; Raja Beharelle et al., 2015; Chakroun et al., 2020; Wiehler et al., 2021), this was not the case for TST data analysed here. This may likely be due to the more complex task structure and resulting higher cognitive demands. Thus, tracking the underlying reward walks and uncertainties associated with them might be too computationally demanding. Participants did, however, show uncertainty-dependent adjustments in their choices in S1 of the task, such that robust exploration bonus parameters > 0 were obtained (see below). Yet, when faced with a more complex task, participants might fall back on simpler, heuristic-based uncertainty proxies to guide their behaviour. In fact, the current results point to an advantage of the more parsimonious counter-based heuristic tested here (M5, see model comparison results in favour of the bandit-heuristic compared to the trial-heuristic model M7). Although a purely temporal indicator (M6 & M7, trial heuristic) may seem more intuitive at first, continuous tracking over the course of up to 300 trials (data1; 200 trials in data2) seems cognitively implausible. The bandit heuristic further has the advantage of visual support throughout the task (clearly distinguishable stimulus identities). These external cues, alongside the limited numerical range, ease the demand placed on working memory processes and thus may ultimately “protect” the MB system (Otto et al., 2013; Brown et al., 2022; Collins et al., 2017; Yoo & Collins, 2022; for a diverging view however see Silva et al, 2022). As the instructed goal for participants was to maximize their pay-outs, placing higher priority on the MB system might be advantageous. As Kool and colleagues (2016) have pointed out, such an advantage is however, dependent upon the task version at hand. This finding is supported by the significant positive association of MB control and rewards earned in data1 (adapted version of the TST) and a lack thereof in data2 (classic TST version). These differential associations also show that previously voiced criticisms and proposed alterations (Kool et al., 2016) have successfully been addressed in the adapted task version used in data1. Parameter estimates for the weight on 2nd-stage Q-values were positively associated with overall pay-out in both data sets and were indeed the only significant association in this regard in data2. Thus, relatively lower MB tendencies in data2 may even be seen as goal-directed in a broader sense, as MB control is commonly viewed as more demanding, while in this case not more rewarding and ultimately too costly.

Taking a closer look at the exploration bonus from the winning model, we show that here too, results from data1 and data2 differ in regard to their association with MB control. Table 7 provides rough estimates of the relative proportion of this bonus in relation to MB control. As pointed out above, MB control was attenuated in data2 compared to data1. Considering the similar posterior estimates for the exploration parameter these proportional differences (V_{EB} / Q_{MB} : 5% vs 17% of in data1 and data2, respectively, Table 7) come to little surprise. Nonetheless, one may speculate that they are not fully independent of aforementioned diverging utility and contribution of MB control. Varying reward magnitudes in the adapted TST version (data1) are likely easier to track than binary outcomes in data2 and may have provided participants with a sufficiently useful and not overly costly reference point for S1 preferences. Tracking underlying reward dynamics based on binary outcomes (data2) on the other hand might be more difficult. The additional heuristic-based uncertainty estimates may have thus been of varying utility during decision-making, depending on the environment (i.e. task version) at hand. These considerations are somewhat speculative in nature and warrant further, more detailed investigations in the future.

The assessment within a laboratory setting further enabled more control over participants' understanding of the task at hand compared to the online sample from which we derived data2. As several scholars have pointed out over the past years, general task understanding and diverging instructions of the TST can have a significant influence on the relative employment of the MB over MF system (e.g. Akam et al., 2015; Feher da Silva & Hare, 2018; Castro-Rodrigues et al., 2022; Hamroun, Lebreton,& Palminteri, 2022). It should be noted that the authors of the original publication of data2 have taken extensive precautions such as training trials and a comprehension test (for details see Gillan et al., 2016) prior to execution of the TST. Nonetheless, insight into participants' model of the task remains reduced within this online context. Noticeable differences in the compensation and therefore incentive for participation may have further influenced individuals' motivational state with regard to more effort placed on task execution (see e.g. Patzelt et al., 2019).

Dual-System Views and the TST

Beside these external influences on relative MB and MF contributions (i.e. task versions, instructions, incentives etc.) broader criticisms regarding their definition within classic dual-system frameworks has been raised. As outlined previously, indices of MB control likely only depict one possible goal-directed strategy subjects use to complete the TST (recall reduced MB control in data2 vs. data1 in light of its utility for reward maximization, i.e. long-term goal-attainment). Consequently, several alternative strategies that may also utilize a model of the environment, rendering them MB in the literal sense, are not accounted for (Feher da Silva & Hare, 2018; 2020; Toyama et al., 2017;2019). Models employed on the other hand, may be skewed, outright incorrect, or employed in a rigid and habitual way, further complicating a clear-cut interpretation of associated indices (see e.g. Seow et al., 2021; Shahar et al., 2019a). The same holds true for potential additional subprocesses which are not represented in classic dual-system views and formalizations thereof (Collins et al., 2017; Collins & Cockburn, 2020; Feher da Silva et al., 2022). At this point it should be noted that several of these issues may also apply to the explore-exploit research and theoretical assumptions works in this field are based on.

To address these concerns several scholars have been developing adapted versions or novel alternatives to these paradigms (see e.g. Kool et al., 2016 and the adapted TST version applied by Mathar et al., 2022; Bruder et al., 2021; Wagner et al., 2021 etc.). To name one prominent example posed as an alternative (or at the least useful supplement) to widely applied classic restless bandit paradigms in the explore-exploit research, Wilson and colleagues (2014) have introduced the *Horizon Task.* This task aimed at the decoupling of reward and information, which are classically confounded and thus hamper the clear distinction between exploration, exploitation and their driving factors. The Horizon Task has been successfully applied in a number of studies and has thus far also undergone several further adaptations (e.g. Feng et al., 2021; Cogliati Dezza et al., 2017; Sadeghiyeh, et al., 2020).

Computational Modelling

Another approach is the development of novel computational models to better delineate the specific processes engaged during task performance. Several researchers have set out to move away from traditional computational accounts, and build upon markedly different conceptualisations of learning and decision processes in this task. One recent example for such efforts comes from Gijsen, Grundei, and Blankenburg (2022): The authors applied an active inference account to TST data and – akin to the procedure laid out here- re-analysed existing data sets. Some of these were in fact better described by the proposed more elaborate models. However, specifically data gathered in an online setting as well as data including a negative reinforcement scheme diverged with regard to model ranking. The preferred model for these data sets (referred to as *online* and *shock* data sets respectively in Gijsen et al., 2022) was in fact more akin to versions tested here. Moreover, it should be noted that the current study followed a different aim in more general terms. As pointed out previously, the TST is widely applied and probably the most utilized paradigm to capture MB and MF behavioural tendencies.

Thus, by extending existing models that are already in use, we hope to balance improving their descriptive ability while at the same time ensuring their applicability for a wide scientific audience. By making the code for the winning model freely available we hope to foster similar efforts (c.f. *Open Code* above).

Limitations of the current study

Common to all computational modelling approaches are basic considerations regarding the limited scope of possible mechanisms accounted for. To put it more bluntly: Results derived from computational models (and their comparison) are ultimately limited to the finite set of processes defined in them. Due to its non-specific applicability this issue may almost seem trivial, but should nonetheless be kept in mind when evaluating and interpreting such results.

In addition to these broader conceptual issues with regard to the TST, more specific limitations of the present study should be noted as well. While results from model comparison clearly favoured all QL over BL models in both data sets it should be noted that the implemented belief updating process in the latter model group depicts an approximation (vs. exact representation) of the true underlying random walk dynamics (control analyses however showed that the empirical reward dynamics closely corresponded to those implemented in all BL models).The Kalman-Filter updating process between trials (Equation 12) in all BL models is analogous to the *forgetting* process implemented in the QL model variants (Equation 4). Thus, for both learning mechanisms we assumed subjective value estimates of unchosen options to move closer to a reasonable estimate (mid-range of possible values). Nonetheless, future applications may consider refining this model aspect.

Despite the fact that the evidence for strategic exploration behaviour was consistently present across both data sets (exploration bonus effects were substantially larger than zero), these effects were numerically small (e.g. mean exploration bonus values ranged between 5% and 17% of MB values). Yet, inclusion of these terms substantially improved model fit. As outlined above, model variants shown in the present study only cover a very limited set of possible exploration mechanisms at play. Different versions of such exploratory behaviour or additional processes not accounted for here are possible, in light of these results plausible or even likely and should be part of future investigations.

These may also include simulations and analyses thereof in order to provide comparison results to those presented here. To our knowledge such analogous simulations from related work are currently not available. Thus, a clear interpretation of the winning model's simulation presented above is hard to reach. Although posterior predictive checks revealed that the winning model (M5) accounted for the over data pattern quite well, it still underpredicted S1 stay-probabilities in both data sets (c.f. Table 4 and Figure 4 above). Future work is required to determine the degree to which this depends on the specific task version employed, or reflects a general shortcoming of current hybrid models.

Another potential limitation is that recently applied DDM choice rules were not used in the present study (Pedersen, Frank, & Biele, 2017). The investigation of reaction time distributions and their relation to information processing and decision-making can provide valuable insights that may complement present results (Shahar et al., 2019b). Parameters derived from models like these have further been linked to various (sub-)clinical symptoms, and thereby also shed light on potential disease mechanisms (see e.g. Forstmann, Ratcliff, & Wagenmakers, 2016; Mandali et al., 2019; Maia, Huys, & Frank, 2017, Sripada & Weigard, 2021). Here however, our goal was first to confirm the existence of the proposed exploration mechanisms in different TST versions. We therefore leave extensions to DDM choice rules to future work.

The empirical data used to develop and test the proposed novel model variants may pose yet another concern. We included two TST versions as a first step, but several additional task variants could be examined in future work in order to further validate the adapted hybrid model. Considering aforementioned ambiguities with regard to the generalizability and transferability of proposed exploration-strategies (and other information-processing steps inherent in the model; e.g. learning mechanisms etc.) these may also be tested for in data from related paradigms posing similar demands (i.e. reward accumulation in light of uncertain, dynamic environments). In order to (at least in part) account for the myriad of contextual factors influencing learning and decision-making in such paradigms, data from within-subject design studies examining specific contextual effects would be required to delineate how the proposed exploration mechanism is modulated by these factors.

An additional issue concerns the generalizability of the present results. To this day, a large part of empirical findings is based on small rather homogenous groups of individuals, namely WEIRD ones (i.e. white, educated, industrialised, rich, & democratic), which also applies to many other data sets in the field. Despite participants' *WEIRDness* , samples are seldomly diverse with regard to age or gender either. In the present case for example, the sample from data1 was exclusively comprised of 18 to 35-year-old heterosexual males (Mathar et al., 2022). While reducing variability in these sample characteristics has its' utility, (improving internal validity and thus enabling more clear-cut interpretations) results are consequently limited to this confined group. Gillan and colleagues (2016) on the other hand employed a more diverse large-scale community sample. Despite lesser concerns regarding diversity, here other limitations that arise due to the online setting and associated factors come into play (e.g. data quality due to false profiles, low incentives, task understanding etc.).

Despite ever growing popularity and application of transdiagnostic as well as dimensional conceptualisations of mental health, a substantial body of research is still based on the comparison of groups defined as either *healthy* or *diseased.* Again, procedures like this have a rational basis, entail advantages, and have produced a wealth of valuable insights. Keeping this and aforementioned progress in mind (Insel et al., 2010; Robbins et al., 2012; Maia & Frank, 2011), it is nonetheless warranted to push further. Future studies that leverage large samples and the natural sub-clinical variation in psychiatric symptomatology these entail, are called for.

CONCLUSION

In summary, we provide computational evidence for a contribution of directed exploration to behaviour in two independent data sets of different variants of the a widely used two-step task (TST). We compared a series of extensions of commonly applied hybrid models for TST behaviour using concepts from exploration-exploitation work (Wilson et al., 2021), and show that a model with a heuristic-based directed exploration term for S1 decisions consistently outperformed both a baseline model without a directed exploration process, and a series of alternative model variants. Future work may extend these approaches to other task variants, and explore the degree to which the observed directed exploration process observed for TST behaviour is sensitive to e.g. individual differences in (sub-)clinical measures of psychopathology.

References

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2016). Computational Psychiatry: Towards a mathematically informed understanding of mental illness. Journal of Neurology, Neurosurgery & Psychiatry, 87(1), 53–63. https://doi.org/10.1136/jnnp-2015-310737

Addicott, M. A., Pearson, J. M., Wilson, J., Platt, M. L., & McClernon, F. J. (2013). Smoking and the bandit: A preliminary study of smoker and nonsmoker differences in exploratory behavior measured with a multiarmed bandit task. Experimental and Clinical Psychopharmacology, 21, 66–73. https://doi.org/10.1037/a0030843

Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L., & Platt, M. L. (2017). A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. Neuropsychopharmacology, 42(10), Article 10. https://doi.org/10.1038/npp.2017.108

Akam, T., Costa, R., & Dayan, P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. PLOS Computational Biology, 11(12), e1004648. https://doi.org/10.1371/journal.pcbi.1004648

Balleine, B., O'Doherty, J. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacol 35, 48–69 (2010). https://doi.org/10.1038/npp.2009.131

Brown, V. M., Chen, J., Gillan, C. M., & Price, R. B. (2020). Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 5(6), 601–609. https://doi.org/10.1016/j.bpsc.2019.12.019

Brown, V. M., Hallquist, M. N., Frank, M. J., & Dombrovski, A. Y. (2022). Humans adaptively resolve the explore-exploit dilemma under cognitive constraints: Evidence from a multi-armed bandit task. Cognition, 229, 105233. https://doi.org/10.1016/j.cognition.2022.105233

Bruder, L. R., Wagner, B., Mathar, D., & Peters, J. (2021). Increased temporal discounting and reduced model-based control in problem gambling are not substantially modulated by exposure to virtual gambling environments (p. 2021.09.16.459889). bioRxiv. https://doi.org/10.1101/2021.09.16.459889

Castro-Rodrigues, P., Akam, T., Snorasson, I., Camacho, M., Paixão, V., Maia, A., Barahona-Corrêa, J. B., Dayan, P., Simpson, H. B., Costa, R. M., & Oliveira-Maia, A. J. (2022). Explicit knowledge of task structure is a primary determinant of human model-based action. Nature Human Behaviour, 6(8), Article 8. https://doi.org/10.1038/s41562-022-01346-2

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. https://doi.org/10.7554/eLife.51260

Cogliati Dezza, I., Yu, A. J., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. Scientific Reports, 7(1), Article 1. https://doi.org/10.1038/s41598-017-17237-w

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. Biological Psychiatry, 82(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. Nature Reviews Neuroscience, 21(10), Article 10. https://doi.org/10.1038/s41583-020-0355-6

Conway, C. C., & Krueger, R. F. (2021). Rethinking the Diagnosis of Mental Disorders: Data-Driven Psychological Dimensions, Not Categories, as a Framework for Mental-Health Research, Treatment, and Training. Current Directions in Psychological Science, 30(2), 151–158. https://doi.org/10.1177/0963721421990353

Dalgleish, T., Black, M., Johnston, D., & Bevan, A. (2020). Transdiagnostic approaches to mental health problems: Current status and future directions. Journal of Consulting and Clinical Psychology, 88(3), 179. https://doi.org/10.1037/ccp0000482

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron, 69(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., & O'Doherty, J. P. (2014). Multiple Systems for Value Learning. In Neuroeconomics (pp. 393–410). Elsevier. https://doi.org/10.1016/B978-0-12-416008-8.00021-8

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. https://doi.org/10.1038/nature04766

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. Neuron, 80(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doody, M., Van Swieten, M. M. H., & Manohar, S. G. (2022). Model-based learning retrospectively updates model-free values. Scientific Reports, 12(1), Article 1. https://doi.org/10.1038/s41598-022-05567-3

Feher da Silva, C., & Hare, T. A. (2018). A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. PLoS ONE, 13(4), e0195328. https://doi.org/10.1371/journal.pone.0195328

Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. Nature Human Behaviour, 4(10), 1053–1066. https://doi.org/10.1038/s41562-020-0905-y

Feher da Silva, C., Lombardi, G., Edelson, M., & Hare, T. A. (2022). A new take on model-based and model-free influences on mental effort and striatal prediction errors (p. 2022.11.04.515162). bioRxiv. https://doi.org/10.1101/2022.11.04.515162

Feng, S. F., Wang, S., Zarnescu, S., & Wilson, R. C. (2021). The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. Scientific Reports, 11(1), Article 1. https://doi.org/10.1038/s41598-021-82530-8

Ferrante, M., & Gordon, J. A. (2021). Computational phenotyping and longitudinal dynamics to inform clinical decision-making in psychiatry. *Neuropsychopharmacology*, *46*(1), Article 1. https://doi.org/10.1038/s41386-020-00852-z

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. Annual Review of Psychology, 67, 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Gershman, S. J., & Daw, N. D. (2012). Perception, action and utility: The tangled skein. Principles of Brain Dynamics: Global State Interactions, 293–312.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. Cognition, 173, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Gershman, S. J. (2019). Uncertainty and exploration. Decision, 6(3), 277–286. https://doi.org/10.1037/dec0000101

Gijsen, S., Grundei, M., & Blankenburg, F. (2022). Active inference and the two-step task. Scientific Reports, 12(1), Article 1. https://doi.org/10.1038/s41598-022-21766-4

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. ELife, 5, e11305. https://doi.org/10.7554/eLife.11305

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive–compulsive disorder. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1655), 20130475. https://doi.org/10.1098/rstb.2013.0475

Goschke, T. (2014). Dysfunctions of decision-making and cognitive control as transdiagnostic mechanisms of mental disorders: Advances, gaps, and needs in current research. International Journal of Methods in Psychiatric Research, 23(S1), 41–57. https://doi.org/10.1002/mpr.1410

Hamroun, S., Lebreton, M., & Palminteri, S. (2022). Dissociation between task structure learning and performance in human model-based reinforcement learning. PsyArXiv. https://doi.org/10.31234/osf.io/2uw85

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. Neuropsychopharmacology, 46(1), Article 1. https://doi.org/10.1038/s41386-020-0746-4

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience, 19(3), 404–413. https://doi.org/10.1038/nn.4238

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. American Journal of Psychiatry, 171(4), 395–397. https://doi.org/10.1176/appi.ajp.2014.14020138

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. American Journal of Psychiatry, 167(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When Does Model-Based Control Pay Off? PLOS Computational Biology, 12(8), e1005090. https://doi.org/10.1371/journal.pcbi.1005090

Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Chapter 7—Competition and Cooperation Between Multiple Reinforcement Learning Systems. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), Goal-Directed Decision Making (pp. 153–178). Academic Press. https://doi.org/10.1016/B978-0-12-812098-9.00007-3

Maia, T. V., Huys, Q. J. M., & Frank, M. J. (2017). Theory-Based Computational Psychiatry. Biological Psychiatry, 82(6), 382–384. https://doi.org/10.1016/j.biopsych.2017.07.016

Mandali, A., Weidacker, K., Kim, S.-G., & Voon, V. (2019). The ease and sureness of a decision: Evidence accumulation of conflict and uncertainty. Brain, 142(5), 1471–1482. https://doi.org/10.1093/brain/awz013

Mathar, D., Wiebe, A., Tuzsus, D., & Peters, J. (2022). Erotic cue exposure increases physiological arousal, biases choices towards immediate rewards and attenuates model-based reinforcement learning. https://doi.org/10.1101/2022.09.04.506507

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. Psychological Review, 126(2), 292–311. https://doi.org/10.1037/rev0000120

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. Trends in Cognitive Sciences, 16(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Moran, R., Keramati, M., Dayan, P., & Dolan, R. J. (2019). Retrospective model-based inference guides model-free credit assignment. Nature Communications, 10(1), Article 1. https://doi.org/10.1038/s41467-019-08662-8

Morris, L. S., Baek, K., Kundu, P., Harrison, N. A., Frank, M. J., & Voon, V. (2016). Biases in the Explore–Exploit Tradeoff in Addictions: The Role of Avoidance of Uncertainty. Neuropsychopharmacology, 41(4), Article 4. https://doi.org/10.1038/npp.2015.208

Moutoussis, M., Eldar, E., & Dolan, R. J. (2017). Building a New Field of Computational Psychiatry. Biological Psychiatry, 82(6), 388–390. https://doi.org/10.1016/j.biopsych.2016.10.007

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013a). The Curse of Planning. Psychological Science, 24(5), 751–761. https://doi.org/10.1177/0956797612463080

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013b). Working-memory capacity protects model-based learning from stress. Proceedings of the National Academy of Sciences, 110(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). Incentives Boost Model-Based Control Across a Range of Severity on Several Psychiatric Constructs. Biological Psychiatry, 85(5), 425–433. https://doi.org/10.1016/j.biopsych.2018.06.018

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, *24*(4), 1234–1251. https://doi.org/10.3758/s13423-016-1199-y

R Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raja Beharelle, A., Polanía, R., Hare, T. A., & Ruff, C. C. (2015). Transcranial Stimulation over Frontopolar Cortex Elucidates the Choice Attributes and Neural Mechanisms Used to Resolve Exploration-Exploitation Trade-Offs. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 35(43), 14544–14556. https://doi.org/10.1523/JNEUROSCI.2322-15.2015

Reiter, A. M. F., Heinze, H.-J., Schlagenhauf, F., & Deserno, L. (2017). Impaired Flexible Reward-Based Decision-Making in Binge Eating Disorder: Evidence from Computational Modeling and Functional Neuroimaging. Neuropsychopharmacology, 42(3), Article 3. https://doi.org/10.1038/npp.2016.95

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. Trends in Cognitive Sciences, 16(1), 81–91. https://doi.org/10.1016/j.tics.2011.11.009

Sadeghiyeh, H., Wang, S., Alberhasky, M. R., Kyllo, H. M., Shenhav, A., & Wilson, R. C. (2020). Temporal discounting correlates with directed exploration but not with random exploration. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-60576-4

Seow, T. X. F., Benoit, E., Dempsey, C., Jennings, M., Maxwell, A., O'Connell, R., & Gillan, C. M. (2021). Model-Based Planning Deficits in Compulsivity Are Linked to Faulty Neural Representations of Task Structure. The Journal of Neuroscience, 41(30), 6539. https://doi.org/10.1523/JNEUROSCI.0031-21.2021

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., NSPN Consortium, & Dolan, R. J. (2019a). Credit assignment to state-independent task representations and its relationship with model-based decision making. Proceedings of the National Academy of Sciences of the United States of America, 116(32), 15871–15876. https://doi.org/10.1073/pnas.1821647116

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., & Dolan, R. J. (2019b). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Computational Biology*, *15*(2), e1006803. https://doi.org/10.1371/journal.pcbi.1006803

Silva, C. F. da, Lombardi, G., Edelson, M., & Hare, T. A. (2022). A new take on model-based and model-free influences on mental effort and striatal prediction errors (p. 2022.11.04.515162). bioRxiv. https://doi.org/10.1101/2022.11.04.515162

Speekenbrink, M. (2022). Chasing Unknown Bandits: Uncertainty Guidance in Learning and Decision Making*. Current Directions in Psychological Science, 31(5*), 419–427. https://doi.org/10.1177/09637214221105051

Sripada, C., & Weigard, A. (2021). Impaired Evidence Accumulation as a Transdiagnostic Vulnerability Factor in Psychopathology. Frontiers in Psychiatry, 12. https://doi.org/10.3389/fpsyt.2021.627179

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Toyama, A., Katahira, K., & Ohira, H. (2017). A simple computational algorithm of model-based choice preference. Cognitive, Affective, & Behavioral Neuroscience, 17(4), 764–783. https://doi.org/10.3758/s13415-017-0511-2

Toyama, A., Katahira, K., & Ohira, H. (2019). Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. Journal of Mathematical Psychology, 91, 88–102. https://doi.org/10.1016/j.jmp.2019.03.007

Vehtari A, Gabry J, Magnusson M, Yao Y, Bürkner P, Paananen T, Gelman A (2023). “loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.” R package version 2.6.0, https://mc-stan.org/loo/.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R. N., Gillan, C., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., & Bullmore, E. T. (2015). Disorders of compulsivity: A common bias towards learning habits. Molecular Psychiatry, 20(3), Article 3. https://doi.org/10.1038/mp.2014.44

Voon, V., Reiter, A., Sebold, M., & Groman, S. (2017). Model-Based Control in Dimensional Psychiatry. Biological Psychiatry, 82(6), 391–400. https://doi.org/10.1016/j.biopsych.2017.04.006

Wiehler, A., Chakroun, K., & Peters, J. (2021). Attenuated Directed Exploration during Reinforcement Learning in Gambling Disorder. The Journal of neuroscience : the official journal of the Society for Neuroscience, 41(11), 2512–2522. https://doi.org/10.1523/JNEUROSCI.1607-20.2021

Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-17977-w

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. ELife, 8, e49547. https://doi.org/10.7554/eLife.49547

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. Current Opinion in Behavioral Sciences, 38, 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001

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), 2074–2081. https://doi.org/10.1037/a0038199

Wyckmans, F., Otto, A. R., Sebold, M., Daw, N., Bechara, A., Saeremans, M., Kornreich, C., Chatard, A., Jaafari, N., & Noël, X. (2019). Reduced model-based decision-making in gambling disorder. Scientific Reports, 9(1), 19625. https://doi.org/10.1038/s41598-019-56161-z

Yip, S. W., Barch, D. M., Chase, H. W., Flagel, S., Huys, Q. J., Konova, A. B., Montague, R., & Paulus, M. (2022). From computation to clinic. Biological Psychiatry Global Open Science.

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. Journal of Cognitive Neuroscience, 34(4), 551–568. https://doi.org/10.1162/jocn_a_01808

SUPPLEMENT

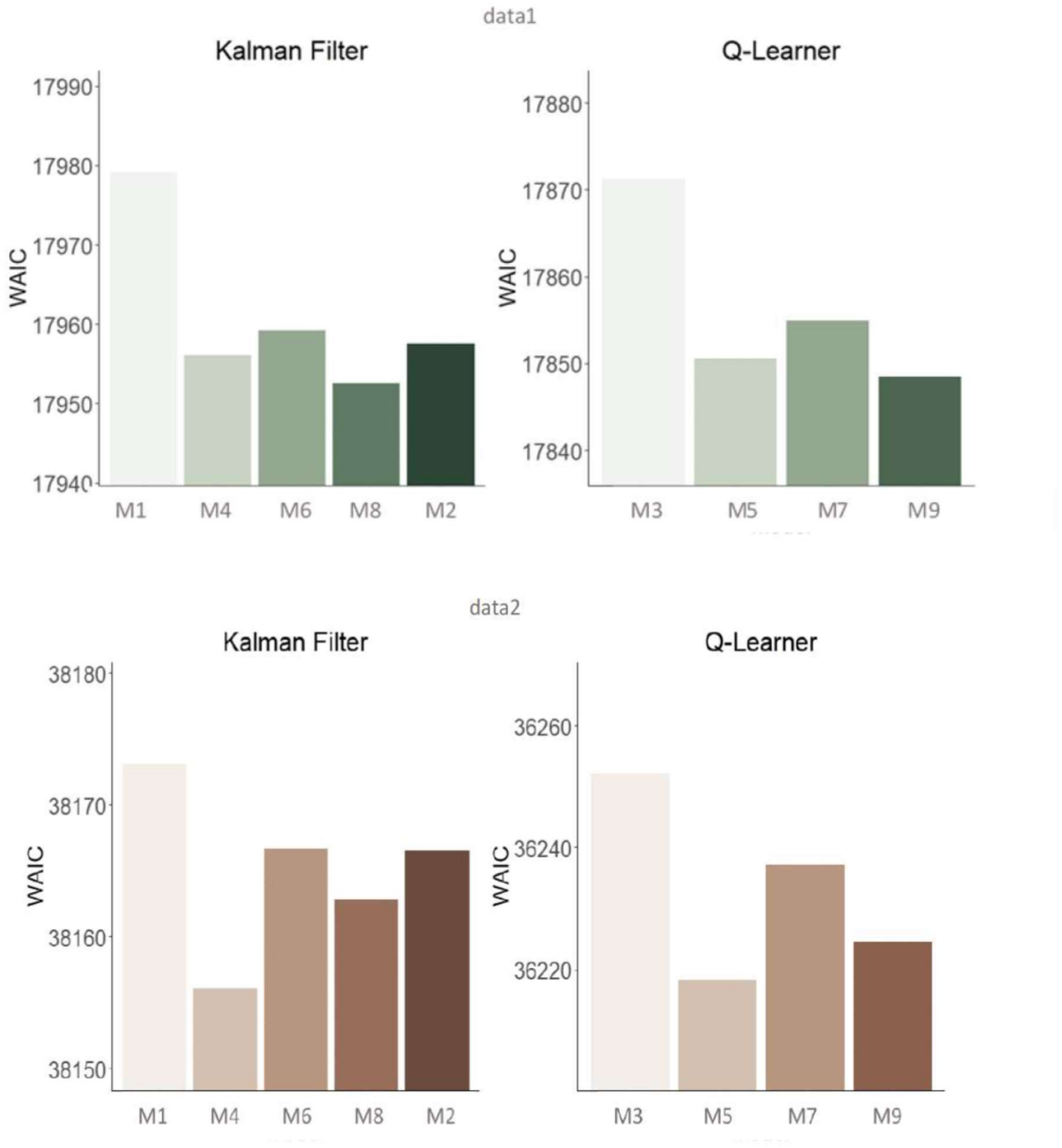
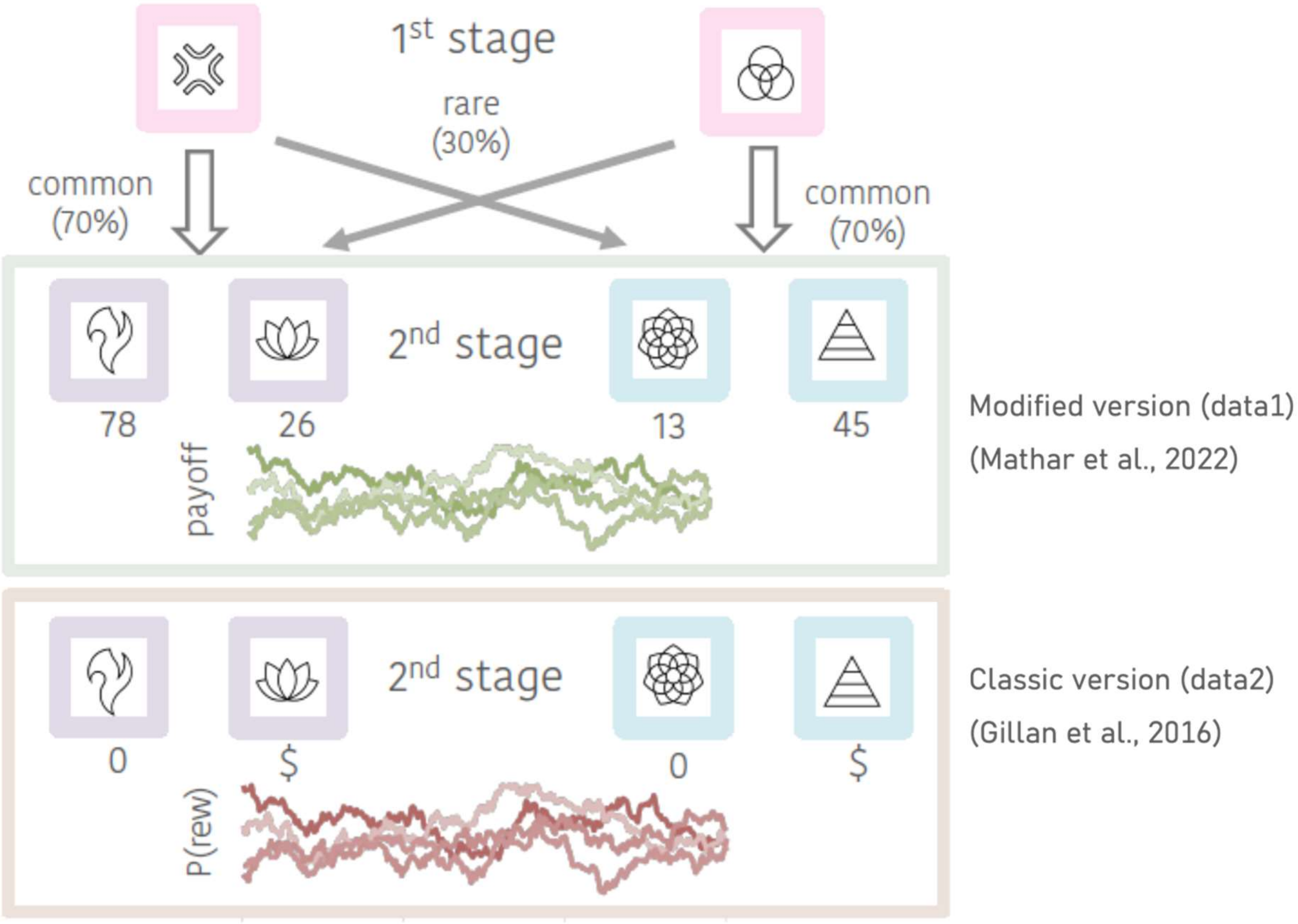
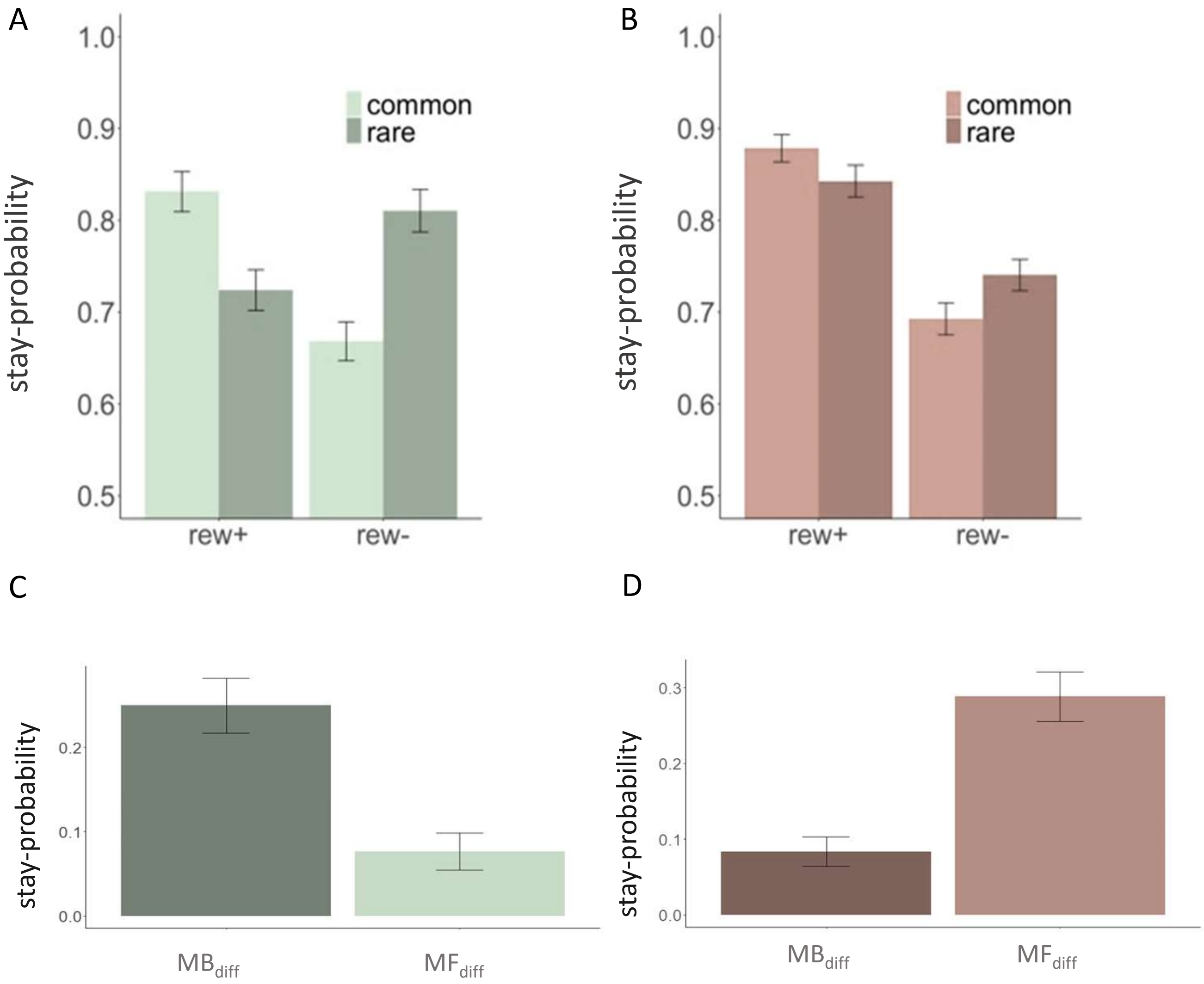
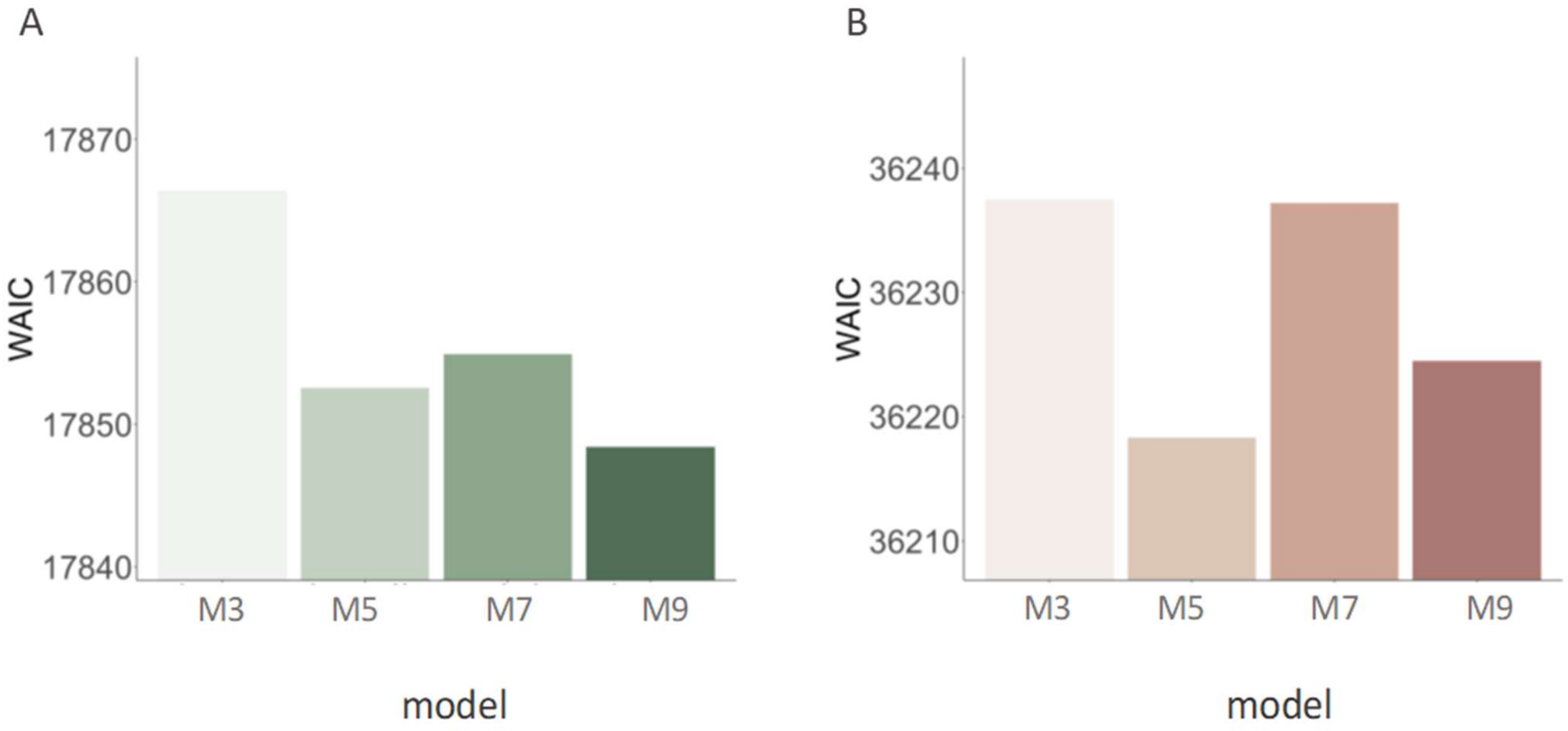
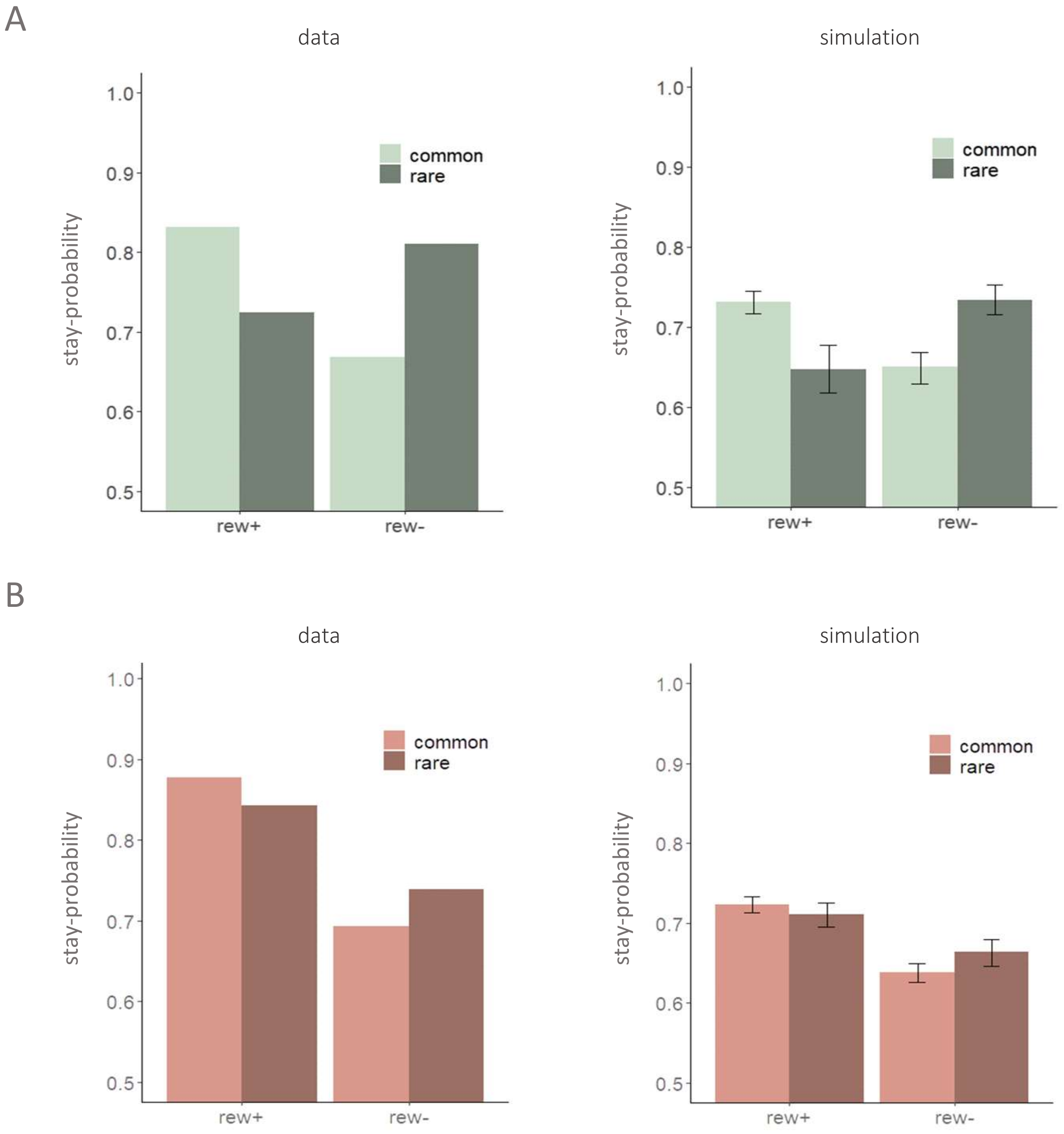


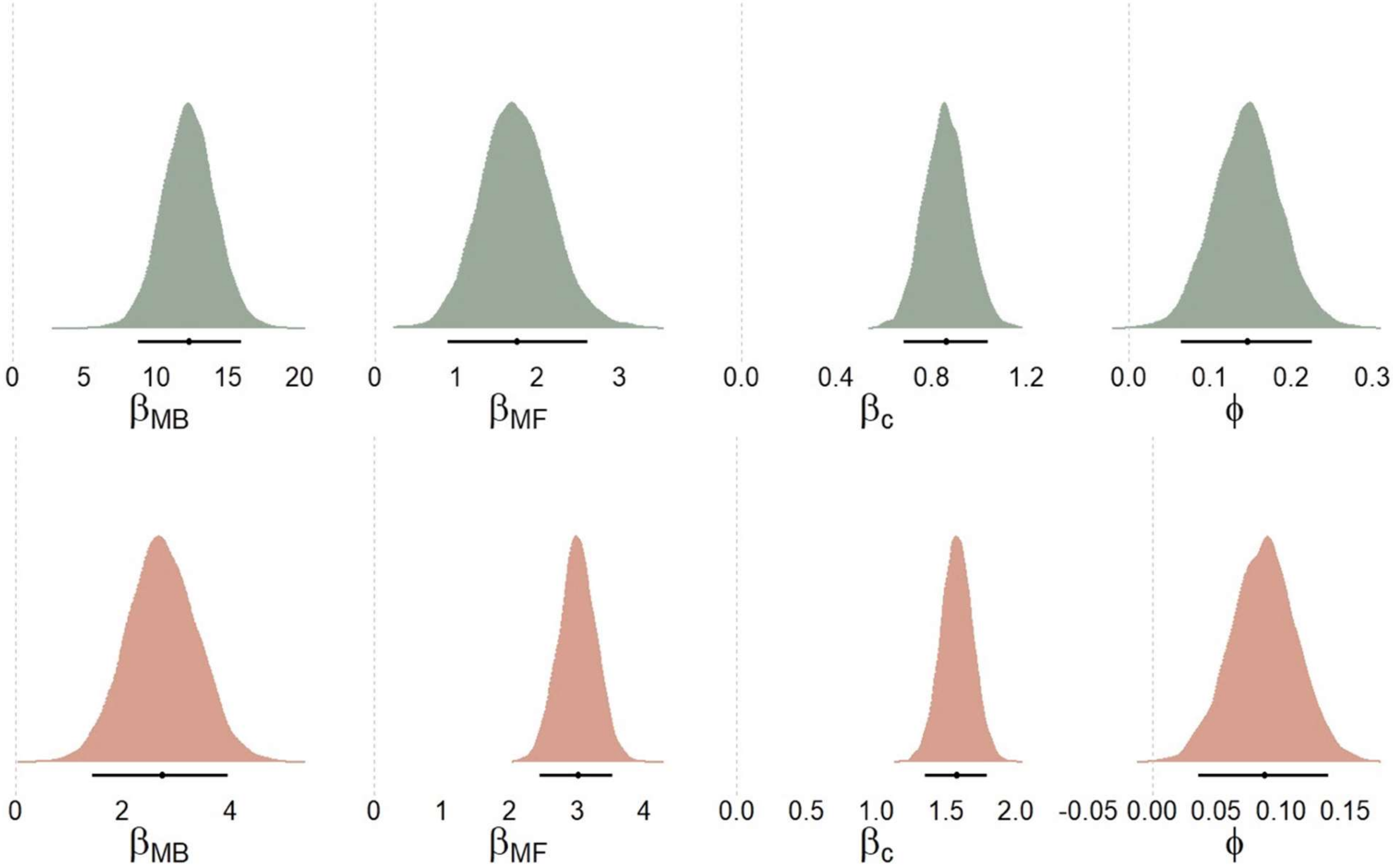
Figure S1. Model Comparison Results for all models considered via the Widely Applied Information Criterion (WAIC). Panel A/B green/orange bar plots refer to data1 and data2, respectively. Kalman-Filter: Bayesian Learner (BL) models.

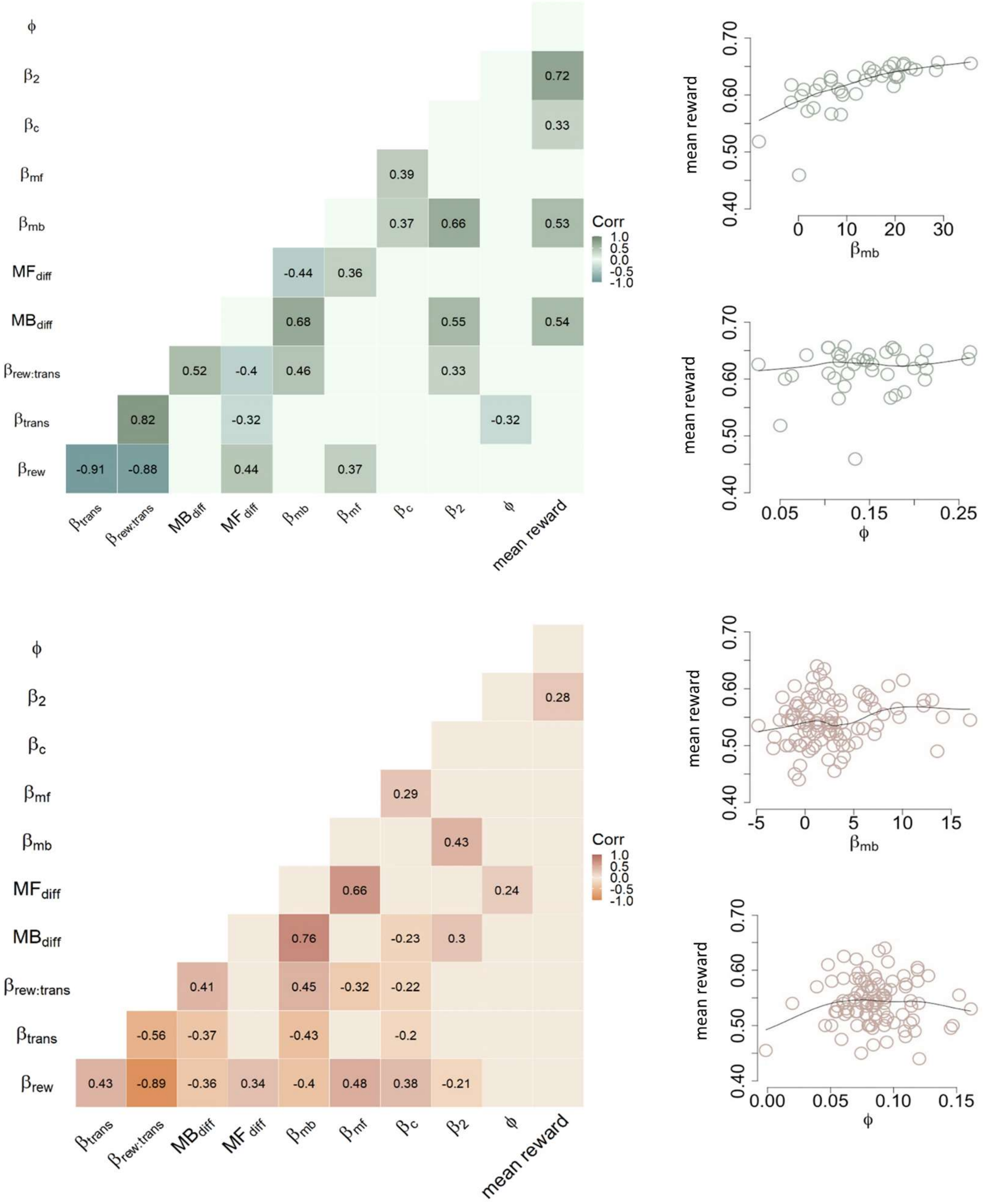

1st stage
rare
(30%)
common
(70%)
common
(70%)
2nd stage
78
26
13
45
payoff
Modified version (data1)
(Mathar et al., 2022)
2nd stage
0
$
0
$
P(rew)
Classic version (data2)
(Gillan et al., 2016)

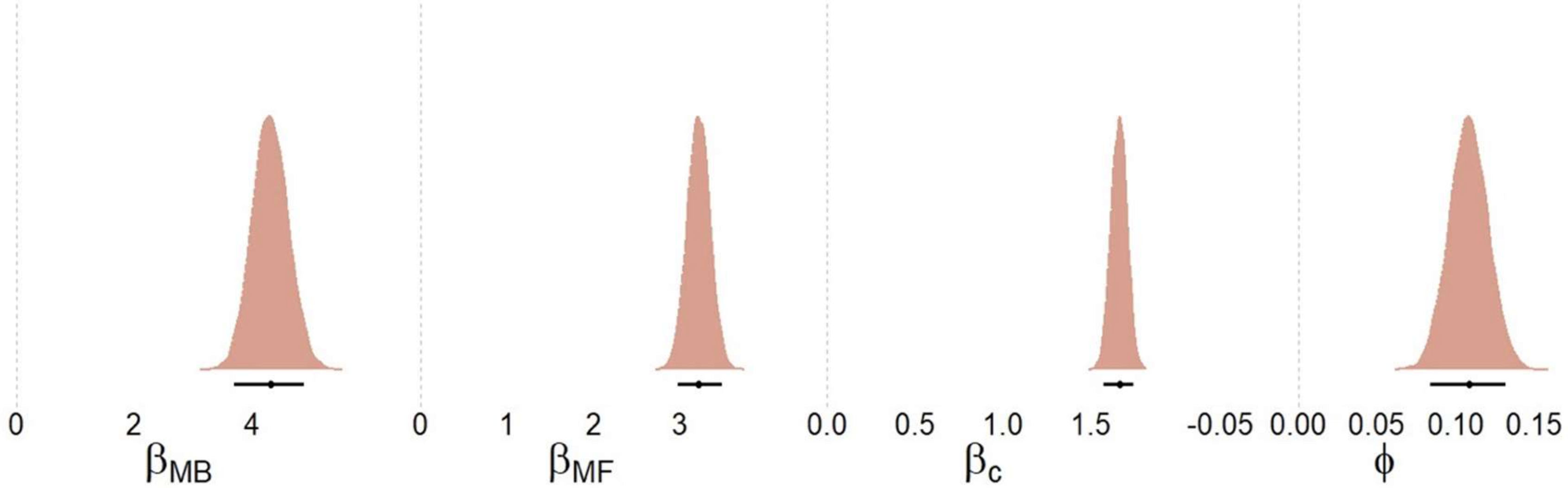

A
1.0
0.9
0.8
0.7
0.6
0.5
stay-probability
common
rare
rew+
rew-
B
1.0
0.9
0.8
0.7
0.6
0.5
stay-probability
common
rare
rew+
rew-
C
0.2
0.1
0.0
stay-probability
MB$_{diff}$
MF$_{diff}$
D
0.3
0.2
0.1
0.0
stay-probability
MB$_{diff}$
MF$_{diff}$

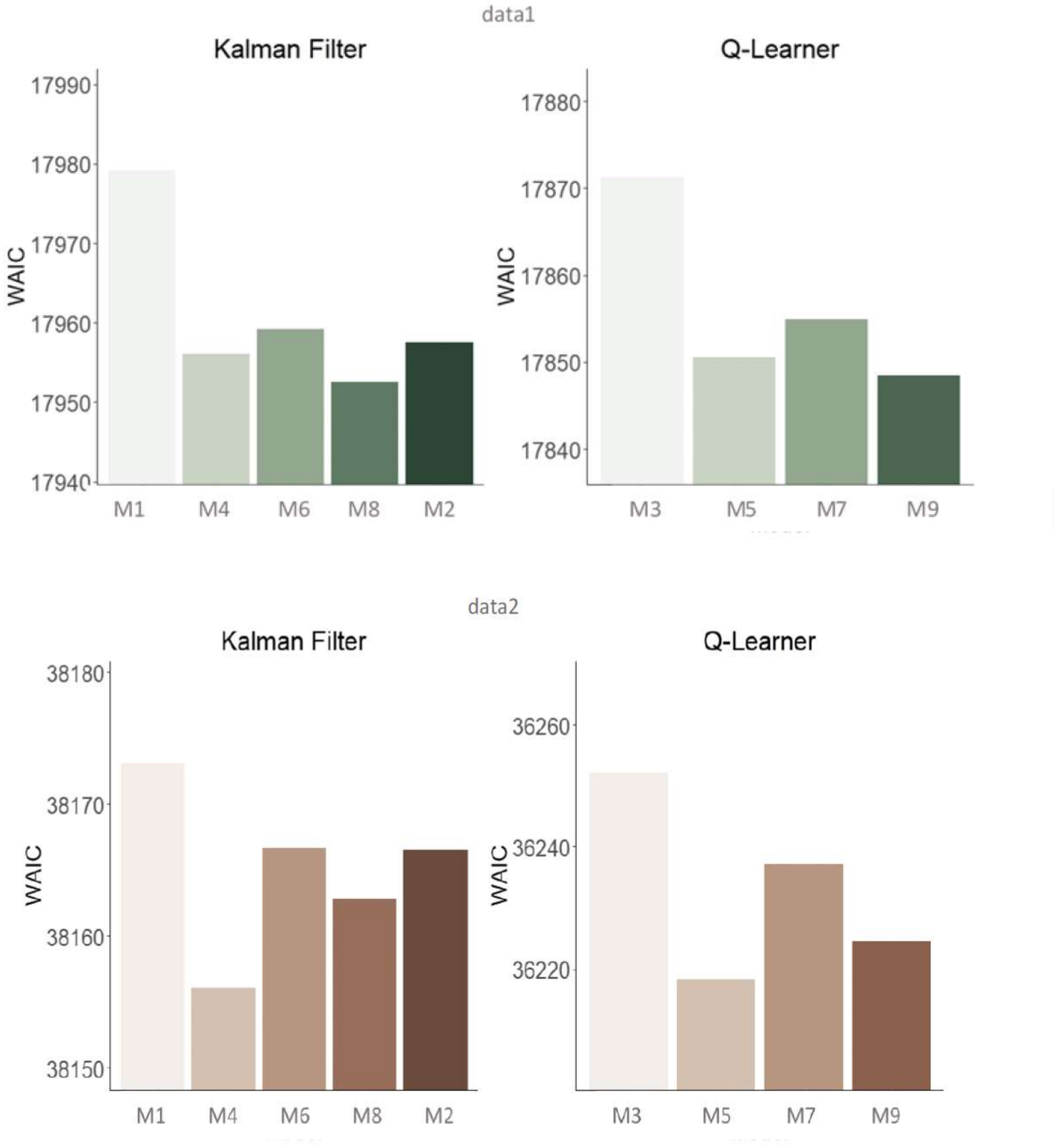

A
WAIC
17870
17860
17850
17840
M3
M5
M7
M9
model
B
WAIC
36240
36230
36220
36210
M3
M5
M7
M9
model


A
data
simulation
stay-probability
1.0
0.9
0.8
0.7
0.6
0.5
common
rare
rew+
rew-
B
data
simulation
stay-probability
1.0
0.9
0.8
0.7
0.6
0.5
common
rare
rew+
rew-


0
5
10
15
20
β_{MB}
0
1
2
3
β_{MF}
0.0
0.4
0.8
1.2
β_c
0.0
0.1
0.2
0.3
ϕ
0
2
4
β_{MB}
0
1
2
3
4
β_{MF}
0.0
0.5
1.0
1.5
2.0
β_c
-0.05
0.00
0.05
0.10
0.15
ϕ


ϕ
β2
βc
βmf
βmb
MFdiff
MBdiff
βrew:trans
βtrans
βrew
0.72
0.33
0.39
0.37
0.66
0.53
-0.44
0.36
0.68
0.55
0.54
0.52
-0.4
0.46
0.33
0.82
-0.32
-0.32
-0.91
-0.88
0.44
0.37
Corr
1.0
0.5
0.0
-0.5
-1.0
βtrans
βrew:trans
MBdiff
MF diff
βmb
βmf
βc
β2
ϕ
mean reward
mean reward
0.40
0.50
0.60
0.70
0
10
20
30
βmb
mean reward
0.40
0.50
0.60
0.70
0.05
0.15
0.25
ϕ
ϕ
β2
βc
βmf
βmb
MFdiff
MBdiff
βrew:trans
βtrans
βrew
0.28
0.29
0.43
0.66
0.24
0.76
-0.23
0.3
0.41
0.45
-0.32
-0.22
-0.56
-0.37
-0.43
-0.2
0.43
-0.89
-0.36
0.34
-0.4
0.48
0.38
-0.21
Corr
1.0
0.5
0.0
-0.5
-1.0
βtrans
βrew:trans
MBdiff
MF diff
βmb
βmf
βc
β2
ϕ
mean reward
mean reward
0.40
0.50
0.60
0.70
-5
0
5
10
15
βmb
mean reward
0.40
0.50
0.60
0.70
0.00
0.05
0.10
0.15
ϕ


0
2
4
β_{MB}
0
1
2
3
β_{MF}
0.0
0.5
1.0
1.5
β_c
-0.05
0.00
0.05
0.10
0.15
ϕ


data1
Kalman Filter
WAIC
17990
17980
17970
17960
17950
17940
M1
M4
M6
M8
M2
Q-Learner
WAIC
17880
17870
17860
17850
17840
M3
M5
M7
M9
data2
Kalman Filter
WAIC
38180
38170
38160
38150
M1
M4
M6
M8
M2
Q-Learner
WAIC
36260
36240
36220
M3
M5
M7
M9