computational psychiatry

Signatures of Perseveration and Heuristic-Based Directed Exploration in Two-Step Sequential Decision Task Behaviour

RESEARCH ARTICLE

ANGELA MARIELE BRANDS
DAVID MATHAR
JAN PETERS
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Processes formalized in classic Reinforcement Learning (RL) theory, such as model-based (MB) control, habit formation, and exploration 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 and perseveration mechanisms. In two independent data sets spanning two different variants of the task, an extended hybrid RL model with a higher-order perseveration and heuristic-based exploration mechanism provided the best fit. While a simpler model with complex perseveration only, was equally well equipped to describe the data, we found 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.

CORRESPONDING AUTHOR:
Angela Mariele Brands
Biological Psychology, Department of Psychology, University of Cologne, Germany
a.brands@uni-koeln.de

KEYWORDS:
computational psychiatry; model-based; exploration; higher-order perseveration; habits; two-step task; neurocomputational endophenotypes

TO CITE THIS ARTICLE:
Brands, A. M., Mathar, D., & Peters, J. (2025). Signatures of Perseveration and Heuristic-Based Directed Exploration in Two-Step Sequential Decision Task Behaviour. *Computational Psychiatry*, 9(1), pp. 39–62. DOI: https://doi.org/10.5334/cpsy.101

DOI: 10.5334/cpsy.101

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 constitutes the core criticism of classic, clear-cut categories of mental disorders. This perspective is captured in more recent approaches to mental health, such as *dimensional psychiatry*. In this view, symptoms exist on a spectrum, with sub-clinical 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 often goes hand in hand with this dimensional view but specifically tackles the traditional symptom-based categorisation and thereby partitioning of mental disorders. High rates of comorbidity present a 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; Dalgleish et al., 2020).

Research into the basic computational processes that may go awry in the case of mental disorders provides important groundwork for these approaches (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).

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). Several key processes have emerged as promising computational endophenotypes mapping onto (sub-) clinical variation in symptoms – *model-based* (MB) control, exploration behaviour and perseveration (Goschke 2014; Addicott et al., 2017; Kool et al., 2018).

MB control utilises a model of the world to predict action outcomes and guide behaviour accordingly. It is thought to act in concert with a simpler, *model-free* (MF) system, which selects actions based on past reinforcement (e.g. Balleine & O'Doherty, 2010; Daw et al., 2011; Daw & O'Doherty, 2014). The exploration-exploitation trade-off (Addicott 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). Directed exploration shares some conceptual features with MB control, as they are both assumed to be goal-oriented and 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, simpler heuristic-based exploration strategies may serve as computationally less costly alternatives (Fox et al., 2020). Instead of a precise model of environmental dynamics (see e.g. Kalman Filter models;. Chakroun et al., 2020; Daw et al., 2006; Speekenbrink, 2022), an agent may utilize a simple proxy measure of environmental uncertainty (and therefore of potential information gain). Perseveration, on the other hand, refers to a general tendency for action repetition (or choice stickiness). It is related to both MB control and exploration, as it is by definition linked to reduced exploration and is often thought to be associated with reduced MB control (see e.g. Voon et al., 2015). In a range of computational models, perseveration is modelled as a subjects' propensity to repeat the directly preceding (t–1) action (*first order perseveration*; FOP). Perseveration can also be conceptualised to extend over several trials (Lau & Glimcher, 2005; Gershman 2020; Miller et al., 2019). In *Higher Order Perseveration* (HOP) subjects' previous actions beyond the last trial (t–1) continue to exert an influence on current actions (see e.g. Bornstein & Banavar, 2023; Miller et al., 2019). Notably, these effects are independent of value (Miller et al., 2019), i.e. independent of the

MF system, and may be more closely related to real-world habitual behaviour than FOP (which in turn is assumed to depict basic motoric components of habitual responding; see e.g. Bornstein & Banavar, 2023; Gershman, 2020; Miller et al., 2019).

Alterations in directed exploration and MB control are thought to underlie habitual and/or compulsive behaviours characteristic of several mental disorders (e.g. Voon et al., 2015) spanning schizophrenia (Culberth et al., 2016), substance use disorders (e.g. Addicott et al., 2013; Morris et al., 2016; Reiter et al., 2016; Sebold et al., 2017; Smith et al., 2020), pathological gambling (e.g. Bruder et al., 2021; Wiehler, Chakroun, & Peters., 2021; Wyckmans et al., 2019), eating disorders (Foerde et al., 2021; Reiter et al., 2017), and obsessive-compulsive disorder (OCD; Banca et al., 2023; Brown et al., 2020; Gillan et al., 2011; Gillan & Robbins, 2014). Similar effects were observed for sub-clinical variations in symptom severity (Gillan et al., 2016; Seow et al., 2021). Reduced MB control could 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). Notably, however, maladaptive behaviours in these groups have also been linked to increased perseveration (SUD: e.g. Leeman & Potenza, 2012; schizophrenia: Waford & Lewine, 2010; OCD: Banca et al., 2023; depression and eating disorders: Voon et al., 2015; Waller et al., 2012). Dysregulations in MB control, exploration and perseveration might therefore constitute potential transdiagnostic endophenotypes in computational psychiatry (Addicott et al., 2017). Interestingly, despite the fact that both MB control and directed exploration have been conceptually linked to the goal-directed system, and perseveration to downregulation thereof, they have largely been studied in isolation.

For more than a decade, the *two-step task* (TST, Daw et al., 2011) has been one key paradigm in the study MB and MF contributions to behaviour, and a central instrument in computational psychiatry (e.g. Voon et al., 2017; Patzelt et al., 2019; Dolan & Dayan, 2013). The TST consists of two stages, which each involve a binary choice and are linked in a (stable/fixed) probabilistic manner (Daw et al., 2011; Figure 1). First-stage choices lead to one of two different second stages (S2, Figure 1) with either high (common transition, 70%) or low probability (rare transition, 30%). 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).

Importantly, MB and MF control make different predictions for S1 choices as a function of reward and transition on the previous trial (c.f. e.g. Daw et al., 2011; Otto et al., 2013b). 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).

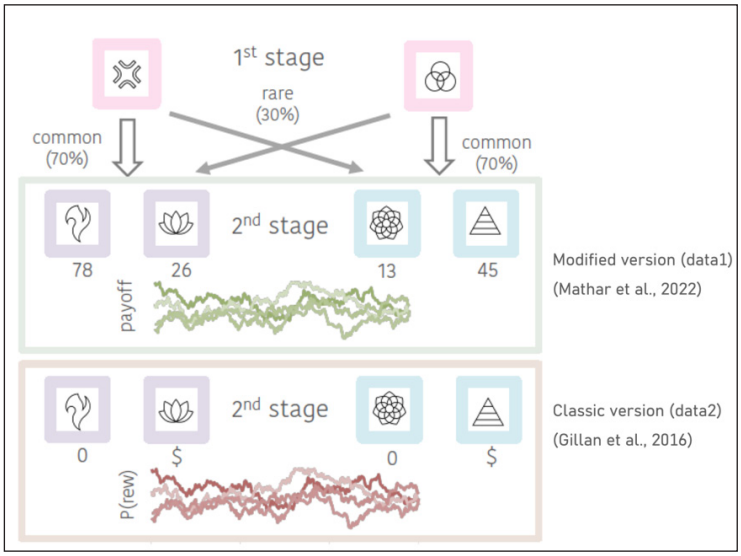

Figure 1 Outline of the two-step task (TST). Transition probabilities from the first stage to the second stage remain the same in both versions of the task. The second stage with a green frame depicts the modified task version employed in data set *data1* (Mathar et al., 2022): after making a S2-choice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

Interestingly, the TST shares several key features with tasks traditionally used to study exploration, such as restless bandit tasks (Daw et al., 2006; Sutton & Barto, 2018; Speekenbrink, 2022; Chakroun et al., 2020; Wiehler et al., 2021). Fluctuating rewards (or reward probabilities, depending on task

DOI: 10.5334/cpsy.101

version) in S2 of the TST afford continuous tracking and introduce potential information gain in S2 as a function of sampling recency of these 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. One way to investigate this possibility 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 followed this approach. Our primary aim was an extension of standard hybrid models of MB and MF control for the TST by incorporating directed exploration strategies. 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). Second, given the conceptual links between perseveration and exploration (see above), we then expanded our model space to additionally examine the role of first-order vs. higher-order perseveration on the TST. 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 of which (including data from clinical groups) are available for re-analysis. Therefore, investigations into additional computational mechanisms that might be reflected in these data could proof valuable for the field of computational psychiatry as well as mathematical psychology in general.

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. The original study covered two conditions, one including an experimental manipulation in the form of erotic pictures which were presented in a block-wise fashion prior to task execution and rated by the participants with regard to their valence and arousal. Data analysed here all stem from the neutral condition, which followed the same procedure, however, only contained neutral images.

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 <150 ms 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 (reward vs. no reward; Figure 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.

DOI: 10.5334/cpsy.101

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 analyse *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 analysing TST behaviour outside a computational modelling 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:

$$\text{pstay} \sim \text{rew} * \text{trans} + (\text{rew} * \text{trans} + 1 \mid \text{subj})$$

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., 2022; Mathar et al., 2022).

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

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

COMPUTATIONAL MODELS

Our final model space consisted of six models in total. Two standard hybrid RL models without exploration terms but with different accounts of perseveration: a common implementation of first-order perseveration (FOP; i.e. repetition of the preceding action) and more temporally extended higher-order perseveration (HOP; c.f. Q + FOP & Q + HOP, respectively). We extended both models with exploration terms that incorporated different ways in which S2 uncertainty might impact first stage choice probabilities (see below). In addition to the model space presented here, we also tested a variety of other alternatives. These included parallel models with a Kalman-Filter learning rule (*Bayesian Learner*) based on computational models regularly applied in the analysis of data from explore-exploit-paradigms (see e.g. Daw et al., 2006; Chakroun et al., 2020). All of these models however, provided a noticeably inferior fit to the data at hand and are thus, not described in more detail here. Further information on the pre-selection process for the final model variants can be found in the Model Comparison section below. Descriptions of the BL model variants are provided in the supplement (c.f. Section Bayesian Learner Models).

Learning Rule

Q-Learner

The learning mechanism and basis for all final model variants is an adaptation of standard hybrid models (e.g. Daw et al., 2011; Otto et al., 2013b). Here, for each S1 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* on trial *t* are updated according to:

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

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

and a constant learning rat 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):

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

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.

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):

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

$$Q_2(s_{2,t}, \bar{a}_{2,t}) = \alpha_3 * Q_2(s_{2,t}, \bar{a}_{2,t}) + (1 - \alpha_3) * 0.5. \quad (5)$$

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

$$P(s_B \mid s_A, a_A) = 0.7, P(s_C \mid s_A, a_B) = 0.7, \text{ or in the alternative case as:}$$
$$P(s_B \mid s_A, a_A) = 0.3, P(s_C \mid s_A, a_B) = 0.3. \quad (6)$$

with

$$P(s_B \mid s_A, a_B) = 1 - P(s_B \mid s_A, a_A) \text{ and } P(s_C \mid s_A, a_A) = 1 - P(s_C \mid s_A, a_B). \quad (7)$$

First-stage Q_{MB} values were then computed as the maximal Q_2 values weighted by their respective transition probabilities. Thus, using the Bellman equation Q_{MB} values are defined as:

$$Q_{MB}(s_A, a_j) = P(s_B \mid s_A, a_j) \max_{a \in \{a_A, a_B\}} Q_2(s_B, a)$$
$$+ P(s_C \mid s_A, a_j) \max_{a \in \{a_A, a_B\}} Q_2(s_C, a) \quad (8)$$

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

$$Q_{MB}(s_{2,t}, a) = Q_{MF}(s_{2,t}, a) = Q_2(s_{2,t}, a). \quad (9)$$

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).

Exploration Bonus

Next to these classic value computations, our extended models also assume a learning and updating process for the informational value of choice options as indicated by the uncertainty associated with them. The following sections describe the different implementations of directed exploration for first stage choices in more detail. 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., 2014; 2021).

We compared two 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. 8). This formalization was based on previous research efforts defining directed exploration as a goal-directed strategy, aiming at long-term reward accumulation via maximal information gain (Wilson et al., 2014; 2021). 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, as formalized in the MB component, transition probabilities are utilized to weigh the uncertainty estimates (vs reward estimates for MB values) associated with S2 options to reflect these assumed deliberate and foresighted aspects.

DOI: 10.5334/cpsy.101

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). In line with the basic idea of a goal-directed exploration mechanism, which however, relies on a simplifies, efficient uncertainty estimate (i.e. counter-heuristic vs. full Bayesian uncertainty tracking) we assume 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 (c.f. Equations 6–8) resulting in:

$$eb_{bandit}(s_A, a_j) = P(s_B \mid s_A, a_j) \sum_a b_{n,t}(s_B, a) + P(s_C \mid s_A, a_j) \sum_a b_{n,t}(s_C, a). \quad (10)$$

This yielded the exploration bonus implemented in variants *QL + BANDIT* and *QL + BANDIT + HOP* (see Equations 16 and 17 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 those 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 *QL+ TRIAL* and *QL + TRIAL + HOP* 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:

$$eb_{trial}(s_A, a_j) = P(s_B \mid s_A, a_j) \sum_a t_{n,t}(s_B, a) + P(s_C \mid s_A, a_j) \sum_a t_{n,t}(s_C, a). \quad (11)$$

In order to match the numerical range to that of the bandit-heuristic described above, counter values for this heuristic were log-transformed. We set the lower bound to 0, so that only non-negative values were considered. Analogous to model Q + BANDIT, action probabilities for first-stage choices were modelled according to Equation 16. The adaptations to yield model Q + BANDIT + HOP were applied here as well, resulting in variant Q + TRIAL + HOP (Equation 17).

Habitual Controller

In addition to the trial-wise updating of Q-values and exploration heuristics, the inclusion of Higher Order Perseveration (HOP) also encompasses continuous tracking and updating. In contrast to the subjective value and uncertainty estimates, here, updates relate to subjects' own history of S1 choices (i.e. are decoupled from reward values and previous S2 actions).

For the QL + HOP model, instead of *rep*(*a*), we included a *habitual controller* (H_t), which accounts for perseveration behaviour in a temporally extended way (Miller et al., 2019). Not dissimilar to previously described TD-learning processes (c.f. Eq. 1 & 2), the habit strength of each choice option is updated according to:

$$H_t = H_{t-1} + \alpha_{HOP} * (\text{rep(a)}_{i,t} - H_{t-1}), \quad (12)$$

DOI: 10.5334/cpsy.101

where rep(a)$_{i,t}$ is the same indicator function used in the basic FOP model variants (c.f. Eq. 13 below). If a S1 choice is repeated on the subsequent *rep*(*a*) equals 1 and 0 otherwise. The parameter α_{HOP} serves as the updating parameter, which determines the extent to which the current choice "overwrites" the previous choice history (i.e. resulting habit strength). This way, the HOP formulation used here includes the FOP variant as a special case for $\alpha_{HOP} = 1$. In contrast, values closer to 0 would result in slower updating of habit strength, and thus, indicate a stronger (i.e. longer lasting) influence of past choices.

Choice Rules

The standard SoftMax function (*SM*, McFadden, 1973; Sutton & Barto, 2018) served as the basis for all choice rules. According to this rule, choice 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, in the QL + FOP model choice probabilities for action *a* at the first stage were modelled as:

$$P(a_{i,t} = a \mid s_{1,t}) = \frac{exp\left[ß_{MB}Q_{MB}(s_{1,t}, a) + ß_{MF}Q_{MF}(s_{1,t}, a) + ß_{persev}rep(a)\right]}{\sum_{a'} exp\left[ß_{MB}Q_{MB}(s_{1,t}, a') + ß_{MF}Q_{MF}(s_{1,t}, a') + ß_{persev}rep(a')\right]}. \quad (13)$$

The parameter $ß_{persev}$ describes the "stickiness" of first stage choices, i.e. FOP. As described above, the indicator function *rep*(*a*) equals 1 if the first-stage choice of the previous trial is repeated and 0 otherwise.

Action probabilities follow the same SoftMax as in Equation 13 above, with the only difference of replacing the indicator function from FOP models with the habit vector H_t, so that:

$$P(a_{i,t} = a \mid s_{1,t}) = \frac{exp\left[ß_{MB}Q_{MB}(s_{1,t}, a) + ß_{MF}Q_{MF}(s_{1,t}, a) + ß_{persev}H_t(a)\right]}{\sum_{a'} exp\left[ß_{MB}Q_{MB}(s_{1,t}, a') + ß_{MF}Q_{MF}(s_{1,t}, a') + ß_{persev}H_t(a')\right]}. \quad (14)$$

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:

$$P(a_{i,t} = a \mid s_{2,t}) = \frac{exp\left[ß_2 Q_2(s_{2,t}, a)\right]}{\sum_{a'} exp\left[ß_2 Q_2(s_{2,t}, a')\right]} \quad (15)$$

The two baseline models QL + FOP and QL + HOP used these basic versions of the SoftMax. These were extended by incorporating terms that account for directed exploration in first stage choices.

Exploration Bonus (eb)

For all FOP variants accounting for directed exploration, *eb* was included in the standard SoftMax and weighted by an additional free parameter ϕ, resulting in the following first-stage choice probabilities (for models including either eb_{bandit} oreb_{trial}):

$$P(a_{i,t} = a \mid s_{1,t}) = \frac{exp\left[ß_{MB}Q_{MB}(s_{1,t}, a) + ß_{MF}Q_{MF}(s_{1,t}, a) + ß_{persev}rep(a) + \phi eb(s_{1,t}, a)\right]}{\sum_{a'} exp\left[ß_{MB}Q_{MB}(s_{1,t}, a') + ß_{MF}Q_{MF}(s_{1,t}, a') + ß_{persev}rep(a') + \phi eb(s_{1,t}, a')\right]}. \quad (16)$$

Note that the analogue HOP models (i.e. Q + BANDIT + HOP & Q + TRIAL + HOP) follow a parallel formalisation of S1 choice probabilities when replacing the indicator function for FOP (*rep*(*a*)) with the habit vector, resulting in:

$$P(a_{i,t} = a \mid s_{1,t}) = \frac{exp\left[ß_{MB}Q_{MB}(s_{1,t}, a) + ß_{MF}Q_{MF}(s_{1,t}, a) + ß_{persev}H_t(a) + \phi eb(s_{1,t}, a)\right]}{\sum_{a'} exp\left[ß_{MB}Q_{MB}(s_{1,t}, a') + ß_{MF}Q_{MF}(s_{1,t}, a') + ß_{persev}H_t(a') + \phi eb(s_{1,t}, a')\right]} \quad (17)$$

DOI: 10.5334/cpsy.101

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 all α^i, and an upper limit of 20 for remaining group-level SD parameters). All learning and decay rates (α_1, α_2, α_3, α_{HOP}) 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).

MODEL	FREE PARAMETERS
Q	α, α_2, α_3, $ß_{MB}$, $ß_{MF}$, $ß_{persev}$, $ß_2$
Q + BANDIT	α, α_2, α_3, $ß_{MB}$, $ß_{MF}$, $ß_{persev}$, $ß_2$, ϕ
Q + TRIAL	α, α_2, α_3, $ß_{MB}$, $ß_{MF}$, $ß_{persev}$, $ß_2$, ϕ
Q + HOP	α, α_2, α_3, α_{HOP}, $ß_{MB}$, $ß_{MF}$, $ß_{persev}$, $ß_2$
Q + BANDIT + HOP	α, α_2, α_3, α_{HOP}, $ß_{MB}$, $ß_{MF}$, $ß_{persev}$, $ß_2$, ϕ
Q + TRIAL + HOP	α, α_2, α_3, α_{HOP}, $ß_{MB}$, $ß_{MF}$, $ß_{persev}$, $ß_2$, ϕ

Table 1 Free and fixed parameters for all models.

Note. Q refers to the basic hybrid model with a FOP term. BANDIT/TRIAL = added first-stage exploration bonus based on respective counter heuristic (c.f. Computational Models section); 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.

All models were implemented using the *STAN* modelling language (version 2.21.0; Stan Development Team, 2020) 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. Using the default settings of the sampling command, initial values for all parameters were randomly drawn from an interval [–2, 2].

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 files 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). Additional information on the alternative model variants not presented here will be made public upon reasonable request.

Open Data

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

Model Comparison

Model comparison was performed using the *loo* package (Vehtari, Gabry, Magnusson, Yao, Bürkner, Paananen &, Gelman, 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 $elpd_{diff}$, the more parsimonious model was chosen. As an additional indicator for model goodness-of-fit point estimates of the *widely applied information criterion* (WAIC) are also reported.

Narrowing the Model Space

Our initial model space also included an alternative Bayesian Learning account which was however, omitted due to relative inferior fit across both data sets. In this first phase of model comparisons, we compared Q-Learner and Bayesian Learner models. Due to the high number of possible combinations if including all proposed extensions, the first round of comparisons only included FOP model variants. As the results clearly favoured the standard Q-Learner (Supplement Figure S1), further HOP extensions (and combinations with exploration components) were performed with this better-suited learning formalisation.

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 behaviour 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 × transition interaction (reflecting MB control).

		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

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

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.

However, regression results along with visual inspection (Table 2, Figure 2) also suggest potential 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). These differences are however, purely descriptive, due to the different study settings and experimental details, which preclude us from directly comparing effects between studies.

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 best-fitting model (which itself thus, always has an $-elpd_{diff}$ of zero). As outlined previously, the Q-Learner models consistently outperformed BL models across both data sets (cf. Figure S1). Models that include a HOP term outperformed all parallel variants accounting for FOP only (Figure 3).

DOI: 10.5334/cpsy.101

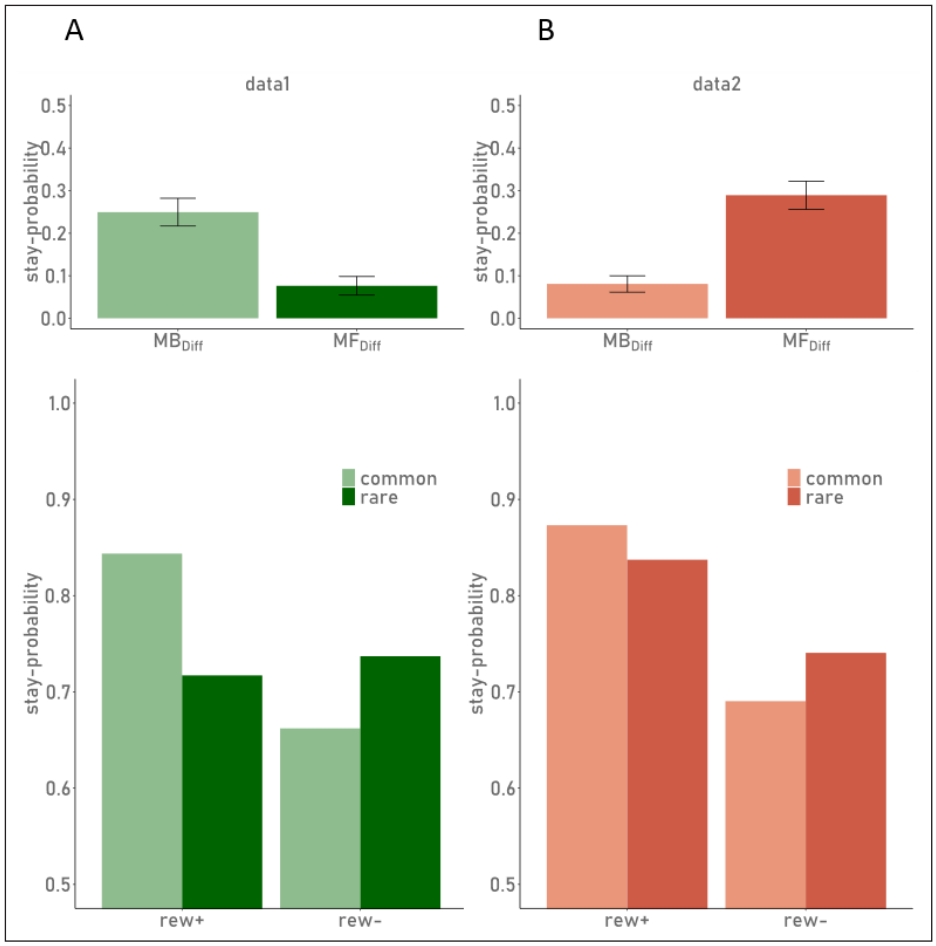

Figure 2 Stay-Probabilities of S1 choices and difference scores. Upper panel: MB and MF difference scores as defined by Eppinger et al. (2013), bar heights depict mean scores over all participants, error bars show the standard error. Lower panel: Probabilities for S1 choice repetition as a function of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

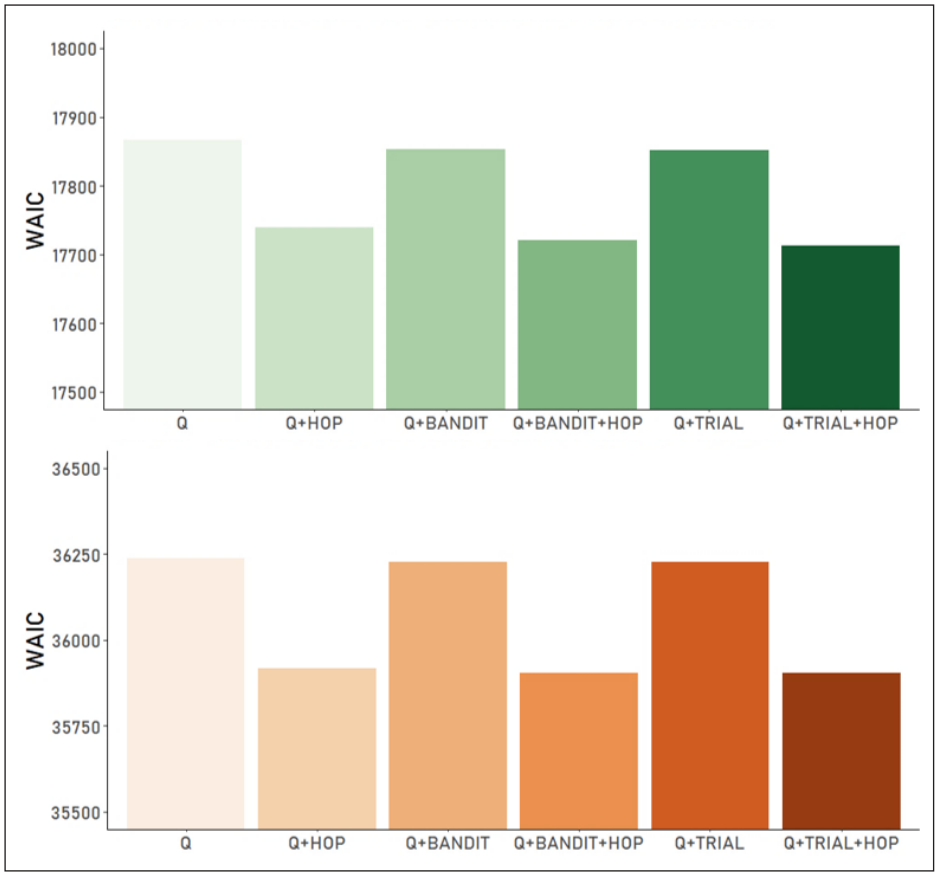

Figure 3 *Model Comparison Results via the Widely Applied Information Criterion (WAIC) for all Q Models (c.f. Table 1).* The upper/lower panel (green/orange bar plots) refer to data1 and data2, respectively. *Bandit/Trial* refer to the model variants with added heuristic-based exploration bonus using stimulus identity/recency, respectively. HOP: model variants with higher order perseveration term; all other versions use a classic FOP term instead (Q, Q+BANDIT, Q+TRIAL).

Within the group of HOP models the variants including an exploration bonus based on a counter heuristic (Q+ BANDIT/TRIAL + HOP) was numerically superior in both data sets (Table 3). However, compared to the Q+HOP model, these differences in goodness of fit (elpd & WAIC) were too small to provide definitive, robust evidence for an added benefit when considering both data sets.

According to some scholars a more complex nested model is warranted if the posterior estimate of an additional parameter in question (i.e. here ϕ) is reliably different from 0 (see e.g. Kruschke, 2011). While this is the case for both data sets (c.f. 95% HDI of ϕ; Table S1), for the main text we are focusing on the reduced model variant (QL+ HOP), based on the inspection of the ratio of benefit in predictive density and its uncertainty (i.e. $elpd_{diff}/se_{diff}$; c.f. Table 3).

Following common conventions (Vethari et al., 2017; Vethari et al., 2023) we thus, interpret the evidence in favour of model Q + TRIAL + HOP (vs. the simpler Q + HOP) in data1 as moderate (28.9/9.6 = 3.01) in data2 however, only as weak (13.8/8.3 = 1.66). Additionally, the more straight-forward interpretation of model parameters as well as our overarching aim of providing a comprehensive and easily applicable refinement of standard hybrid models resulted in the decision for the more parsimonious model version.

Note however, that model fits for data1 were notably improved by inclusion of the exploration term, which was the case regardless of the models' learning component (i.e. also present in BL and FOP variants, see Figure S2). The marked differences in sample, experimental procedure and task version between data sets may likely explain the specificity of superior model fits due to the inclusion of the exploration mechanism. For completion, therefore, all results from the Q+TRIAL+HOP model for Data1 are provided in the supplement, while the main text will focus on the more parsimonious Q+HOP model.

DATA SET	MODEL	$-elpd_{diff}$	se_{diff}	*WAIC*
data1	Q + HOP	−28.9	9.6	17750.21
	Q + BANDIT + HOP	−4.0	6.2	17715.03
	Q + TRIAL + HOP	0.0	0.0	17714.46
data2	Q + HOP	−13.8	8.3	35905.27
	Q + BANDIT + HOP	−11.2	6.2	35887.17
	Q + TRIAL + HOP	0.0	0.0	35871.03

Table 3 Results from model comparison of QL-models with a HOP extension using leave-one-out cross-validation (LOO).

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 Q + TRIAL + HOP), as there is no difference between the best-fitting model and itself, values in the first column are always zero.

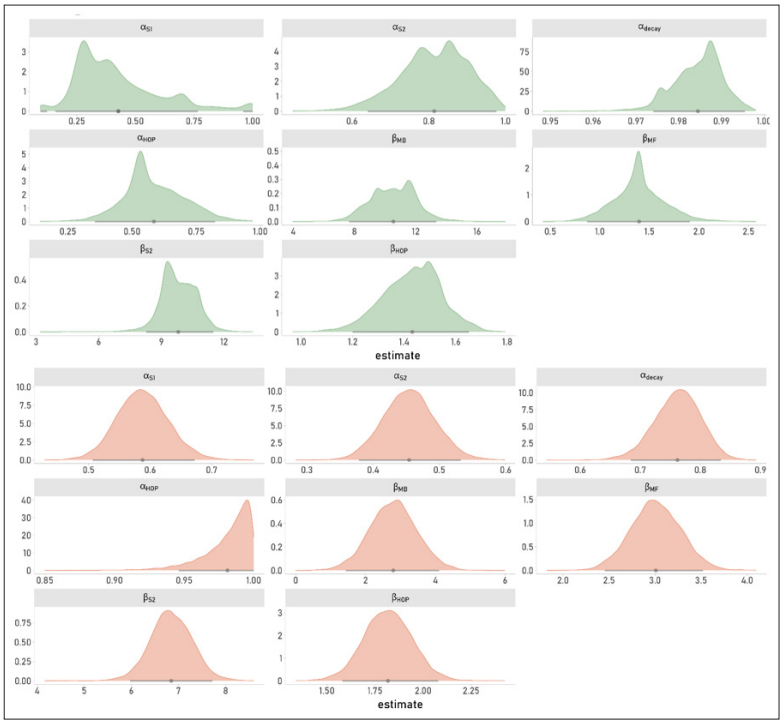

Figure 4 Posterior Distributions of Group-Level Mean Parameters From Model Q + HOP. Solid gray lines show the 95% highest density interval (HDI) and dots depict the point-estimate of the mean. Panels A and B (green and orange plots) show results on the basis of data sets data1 and data2, respectively.

DOI: 10.5334/cpsy.101

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 × 4 chains) 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-behaviour (shown above for empirical data, c.f. Table 2, Figures 2 and 5) for the simulated data sets. Visual inspection (Figure 5) revealed an underestimation of "*stay*"-probabilities (P(stay)) 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 Q +HOP.

Note. Summary statistics are based on the comparison of individuals' choices with model predictions, which were pooled and averaged for each data set.

DATA SET	MIN	25th PERCENTILE	MEDIAN	MEAN	75th PERCENTILE	MAX
data1	.519	.638	.764	.748	.841	.916
data2	.505	.687	.767	.754	.829	.977

Posterior Distributions

Figure 4 shows the posterior distributions of group-level parameters underlying S1 choices in the best-fitting model. The resulting posterior estimates mirrored the model-agnostic analyses portraited above, showing a descriptively more pronounced MB component ($ß_{mb}$) in data1 vs. data2 (Figure 4). The step size parameter α_{HOP}, also exhibits differences between the two data sets. Again, it should be noted however, that such differences are not directly interpretable due to the distinct and independent experimental settings. It is, nonetheless, interesting to note that the step-size parameter in data2 is estimated to be close to 1, which leads the HOP term to approximate FOP (in the case of $\alpha_{HOP} = 1$ only the last step is considered with regard to perseveration). Habit-updating in data1 in contrast, seems to occur in a slower, more gradual fashion, thus, more closely resembling HOP.

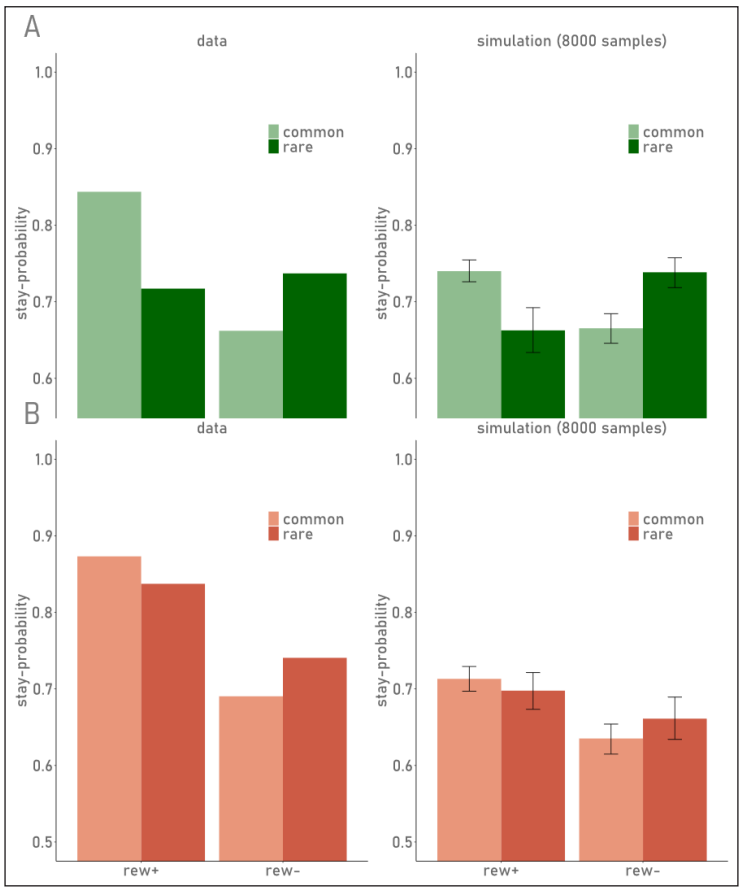

Figure 5 Probabilities of S1 choice repetition as a function of reward and transition type. Y-axis: Stay probabilities for 1st stage choices; data: empirical stay probabilities from data sets data1 (panel A; green) and data2 (panel B; orange). simulation: stay-probabilities from N = 8000 simulated choice sequences per subject, derived from the winning model (Q + HOP).; 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 point-estimates of all hyperparameter means are shown in Table 5, results from the model extension with the directed exploration term based on the *TRIAL* heuristic can be found in the supplement (Table S1, Figure S2). As mentioned above, results from this extended model indicate that meaningful traces of strategic exploratory behaviour were present in both data sets (group-level mean of the exploration bonus parameter (ϕ) was positive in both data sets and the 95% HDI did not overlap with 0, Table S1, Figure S2).

PARAMETER	data1		data2	
	$MEDIAN_x$	95%HDI	$MEDIAN_x$	95%HDI
α_1	0.38	[0.10, 0.83]	0.59	[0.51, 0.67]
α_2	0.82	[0.64, 0.96]	0.45	[0.38, 0.53]
α_3	0.99	[0.97, 1.00]	0.76	[0.68, 0.83]
α_{HOP}	0.57	[0.35, 0.82]	0.98	[0.95, 1.00]
$ß_{mb}$	10.59	[7.65, 13.33]	2.80	[1.44, 4.11]
$ß_{mf}$	1.39	[0.87, 1.91]	3.00	[2.46, 3.52]
$ß_{HOP}$	1.44	[1.20, 1.65]	1.82	[1.58, 2.10]
$ß_2$	9.76	[8.26, 11.49]	6.84	[5.97, 7.72]

Table 5 Posterior Estimates of Group-Level Parameters from Model Q + HOP.

Note. Posterior point-estimates of hyperparameter medians and corresponding 95% highest density intervals (95%HDI) for data1 and data2 from the winning model (Q + HOP) for all subject-level parameters × listed in the first column.

Correspondence with Model-Agnostic Analyses

In order to investigate how parameters derived from the model relate to model-agnostic indices of MB and MF behaviour 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:

$$\text{pstay} \sim \text{rew} * \text{trans} + (\text{rew} * \text{trans} + 1).$$

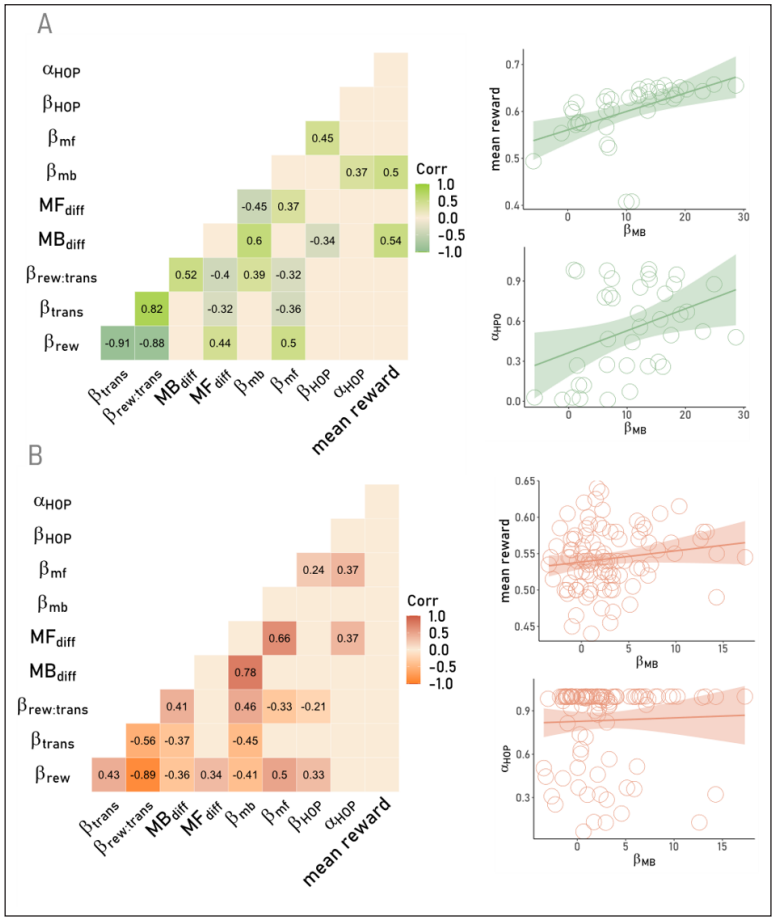

Figure 6 Associations of model-agnostic and model-derived indices of MB and MF control for data1 **(a)** and data2 **(b)**. Empty tiles (left panels) indicate non-significant associations. $ß_{rew}$, $ß_{trans}$, $ß_{rew:trans}$: regression weights for main effects of reward, transition type and their 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; $ß_{HOP}$: S1 higher order perseveration parameter; mean reward: mean reward gained throughout TST (data1: 300 trials, data2: 200 trials). Right panel: association of model-derived MB ($ß_{MB}$) and habit step-size parameter α_{HOP} with mean reward. Circles depict individual participants. Plots in panel A (top row, green) are based on data1, plots in panel B (bottom row, orange) are based on data2.

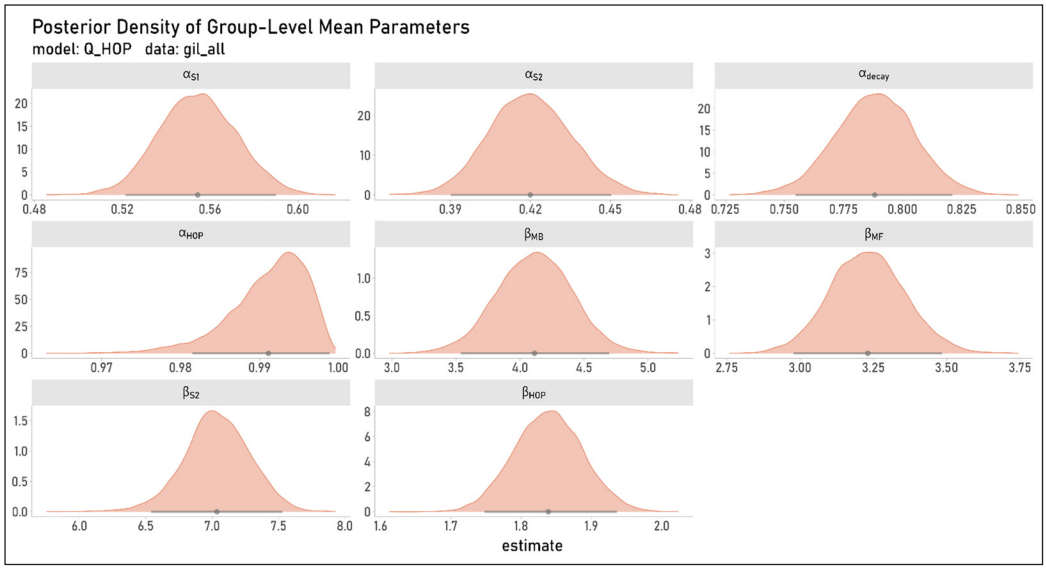

Figure 7 Posterior Density Estimates Based on The Full Sample of Data2. Group-level parameter estimates from model variant Q+HOP derived from fitting the full sample of the original publication (Gillan et al., 2016; N = 548; Experiment 1). The lower panel of Figure 4 shows corresponding results based on data2 (N = 100). Grey dots indicate the mean point-estimate, bars depict the 95%-HDI.

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 MB components ($ß_{MB}$, MB_{diff}) showed a moderate to strong association to the mean overall pay-out 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 6B). The choice weight parameter for the HOP extension ($ß_{HOP}$) was positively associated with the model-derived parameter for MF control across both data sets. In contrast, the second newly introduced parameter, HOP step-size α_{HOP}, showed a diverging pattern. While positively associated with the model-based MB component in data1, in data2 there was no such relation but instead a positive association with the model parameter for MF control. This qualitative pattern may point to a differentiation in parameter relations when considering FOP (c.f. data2 α_{HOP} close to 1) vs. HOP.

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 models Q + HOP and Q + TRIAL + HOP (c.f. Methods described above) using the full sample of experiment 1 (N = 548) from the original publication. Modelling results were comparable to those reported above for the initial subsample (data2). Both models converged equally well (all $\hat{R} < 1.1$). However, in contrast to the ambiguously small numerical advantage for the exploration model over its reduced version in data2 (c.f. Table 3), this benefit was more clearly pronounced in the full sample, where the Q + TRIAL + HOP model outperformed the Q + HOP variant (Supplement Table S3). The posterior estimates of group-level parameters from this model based on the full sample (N = 548) mirrored results reported above (c.f. Figures 4 and 7).

DISCUSSION

Here we extended existing hybrid models of TST behaviour with mechanisms implementing directed (uncertainty-based) exploration and value-free perseveration (higher-order perseveration, HOP). To this end, we considered several ways in which uncertainty may guide participants choices in stage 1 of the task, building upon insights from the exploration-exploitation literature (e.g. 4-armed restless bandit; Daw et al., 2006; Chakroun et al., 2020; Wiehler et al., 2021), as well as first-order vs, higher-order choice repetition effects (References see above). While the TST does not clearly decouple reward and information (in contrast to more specific tasks designed to study exploration, e.g. Wilson et al. 2014), the aim of the present study was to examine potential benefits of incorporating exploration and perseveration mechanisms in models for one of the most widely-used tasks in computational psychiatry.

An uncertainty-dependent learning model (Bayesian learning models using a Kalman Filter) did not provide an advantage over a classic Q-Learner algorithm. We generally observed a positive heuristic-based exploration effect for stage 1 of the task, but the improvement in model fit was only reliable in some analyses (e.g. data1 and full but not reduced sample of data2), and the effect was generally small, such that our primary analyses focused on models without directed exploration terms. In contrast, the inclusion of HOP yielded clear benefits across both independent data sets, which applied two different versions of the TST assessed in different settings (laboratory vs. online sample).

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 qualitative differences in the results obtained from both data sets.

UNCERTAINTY-BASED EXPLORATION ON THE TST

While human choice behaviour in the restless 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 exhibited evidence for uncertainty-dependent exploration effects on S1 choice behaviour. Specifically, exploration bonus parameters were generally reliably >0 across data sets and model variants, dovetailing with previous results using restless bandit tasks (Chakroun et al., 2020; Speekenbrink 2022; Wiehler et al., 2021). At the same time, model comparison in some cases failed to provide conclusive evidence in favour of the inclusion of directed exploration terms (see e.g. the N = 100 subsample for data2), leading to a somewhat inconclusive situation where a nested model parameter was clearly different from zero (indicating a positive effect of directed exploration on S1 choice behaviour), and model comparison was partly inconclusive. Nonetheless, this suggests that the strategic utilisation of uncertainty in TST choice behaviour should be considered in future investigations, while our results also suggest that these effects in TST S1 choice behaviour may be overall consistent, but likely small.

PERSEVERATION BEHAVIOUR ON THE TST

Next to the employment of uncertainty, participants' propensity for perseveration and MB control seem to also afford more nuanced investigation in the context of TST data (and beyond).

The present results show a clear improvement in model fit across both task versions and study set-ups due to the inclusion of S1 HOP behaviour. Descriptively, however, results for data1 and data2 differed. In light of the pivotal role of habits in common theories of addiction (c.f. Everitt & Robbins, 2005; Voon et al., 2017), a more nuanced definition and formalisation is called for (see e.g. Nebe et al., 2024). Repetitive behaviours that are based on e.g. one's own choice history (i.e. HOP) may serve to stabilise behaviour over time. On the other hand, perseveration may also stem from skewed or erroneous value-estimates or lack of behavioural flexibility. Accounting for HOP therefore leads to cleaner process estimates for other mechanisms included in a model.

Though only descriptive in nature at this point, based on the qualitative, diverging pattern in present results, more FOP-like behaviour, as seen in data2 (i.e. step-size close to 1; classic TST version) seems to be associated with MF mechanisms. The analogous, however, more HOP-like behavioural pattern in data1 (i.e. step-size close to 0.5; modified TST) in contrast, showed association with indices of MB control. These findings suggest that the dynamics and associations of decision-making subcomponents may be of interest for our understanding of potential dysregulations along the mental health spectrum. A computational modelling approach as done here, can potentially help to tease apart facets of perseveration.

The potentially differential employment of MB control in data1 and data2 may also be viewed in light of factors specific to each study and the resulting utility of this strategy. As the instructed goal for participants in both studies was to maximize their pay-outs, placing higher priority on the MB system might be seen as clearly 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 (modified 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. Thus, relatively lower MB behaviour 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.

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 along other aspects such as motivation, situational external influences etc. 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., 2022 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*. As alluded to previously, this paradigm is 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 complementary approach is the development of more precise computational models to better delineate the specific processes engaged during task performance. 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 which were better accounted for by the proposed more elaborate models. However, specifically data gathered in an online setting as well as data including a negative reinforcement scheme exhibited differences in 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 arguably the most widely-used paradigm to capture MB and MF behaviour. 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 best-fitting 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. 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. For example, alternative implementations of exploration and/or additional processes not accounted for here could be taken into account in future investigations.

In addition to the 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, the implemented belief updating process in the latter model family is 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). Note that the Kalman-Filter updating process between trials (Supplement Equation 4) in all BL models is analogous to the *forgetting* process implemented in the QL model variants (Equations 4 and 5). 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.

As discussed above, despite the fact that directed exploration estimates were consistently positive across data sets and models, these effects were numerically small. In addition, model comparison (e.g. Q + TRIAL +HOP vs. Q + HOP) did not unequivocally favour the exploration variants when considering both data sets and data2 subsamples. Model Q + TRIAL + HOP yielded a reliably superior fit when considering the full sample (N = 548, c.f. Table S3), the lack of a reliable benefit in the subsample (data2; N = 100) suggests that exploration effects in TST data (in particular the classic TST version assessed in an online study) may be numerically small.

To our knowledge, model simulations analogous to the posterior predictive checks carried out here are currently not available from related work. This complicates the interpretation of our simulation results. Posterior predictive checks revealed that the best-fitting model (Q + HOP) accounted for the overall 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 therefore 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 considered 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). Because our goal was first and foremost to confirm the advantage of proposed model extensions in different TST versions, we leave the application of DDM choice rules to future work.

The empirical data used to develop and test the proposed novel model variants may pose yet another limiting factor. 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- and perseveration-strategies these should clearly be tested for in further, heterogeneous TST data sets. 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 explicitly examining specific contextual effects would be required to delineate how the proposed HOP and exploration mechanisms are modulated by these factors.

An additional issue concerns the generalizability of the present results. 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

Here we compared a series of extensions of commonly applied hybrid models for TST behaviour using concepts from the exploration-exploitation literature and work on perseveration behaviour (Miller et al., 2019; Wilson et al., 2021). Results provide computational evidence for a contribution of higher order perseveration to behaviour in two independent data sets of different variants of the a widely used two-step task (TST). A model with a higher order perseveration term for S1 decisions consistently outperformed standard models accounting only for first-order perseveration. Inclusion of a heuristic-based directed exploration term generally yielded positive exploration bonus parameters, similar to related work using other reinforcement learning tasks. However, these exploration effects were overall small, and model comparison was in some cases inconclusive. Future work may extend these approaches to other task variants, and explore the degree to which directed exploration and/or higher-order perseveration effects in TST behaviour are sensitive to e.g. individual differences in (sub-)clinical psychopathology.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplement.** Additional Modelling Results. DOI: https://doi.org/10.5334/cpsy.101.s1

FUNDING INFORMATION

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

COMPETING INTERESTS

The authors have no competing interests to declare.

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.

AUTHOR AFFILIATIONS

Angela Mariele Brands orcid.org/0000-0003-0422-8160
Biological Psychology, Department of Psychology, University of Cologne, Germany
David Mathar orcid.org/0000-0003-3411-7867
Biological Psychology, Department of Psychology, University of Cologne, Germany
Jan Peters orcid.org/0000-0002-0195-5357
Biological Psychology, Department of Psychology, University of Cologne, Germany

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., 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

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

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

Banca, P., Ruiz, M. H., Gonzalez-Zalba, M. F., Biria, M., Marzuki, A. A., Piercy, T., Sule, A., Fineberg, N. A., & **Robbins, T. W.** (2023). Action-sequence learning, habits and automaticity in obsessive-compulsive disorder. *eLife*, *12*. https://doi.org/10.7554/eLife.87346

Bornstein, A., & **Banavar, N. V.** (2023). Multi-plasticities: Distinguishing context-specific habits from complex perseverations. *PsyArXiv*. https://doi.org/10.31234/osf.io/t7vsc

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

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

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & **Barch, D. M.** (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, *125*, 777–787. https://doi.org/10.1037/abn0000164

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

Eppinger, B., Walter, M., Heekeren, H. R., & **Li, S.-C.** (2013). Of goals and habits: Age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience*, *7*, 253. https://doi.org/10.3389/fnins.2013.00253

Everitt, B. J., & **Robbins, T. W.** (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. https://doi.org/10.1038/nn1579

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

Foerde, K., Daw, N. D., Rufin, T., Walsh, B. T., Shohamy, D., & **Steinglass, J. E.** (2021). Deficient Goal-Directed Control in a Population Characterized by Extreme Goal Pursuit. *Journal of Cognitive Neuroscience*, *33*(3), 463–481. https://doi.org/10.1162/jocn_a_01655

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

Fox, L., Dan, O., Elber-Dorozko, L., & **Loewenstein, Y.** (2020). Exploration: From machines to humans. *Current Opinion in Behavioral Sciences*, *35*, 104–111. https://doi.org/10.1016/j.cobeha.2020.08.004

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

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394. https://doi.org/10.1016/j.cognition.2020.104394

Gershman, S. J., & **Daw, N. D.** (2012). Perception, action and utility: The tangled skein. *Principles of Brain Dynamics: Global State Interactions*, 293–312. https://doi.org/10.7551/mitpress/9108.003.0015

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., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & **de Wit, S.** (2011). Disruption in the Balance Between Goal-Directed Behavior and Habit Learning in Obsessive-Compulsive Disorder. *American Journal of Psychiatry*, *168*(7), 718–726. https://doi.org/10.1176/appi.ajp.2011.10071062

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., 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*. https://doi.org/10.1115/1.3662552

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

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Lau, B., & **Glimcher, P. W.** (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579. https://doi.org/10.1901/jeab.2005.110-04

Leeman, R. F., & **Potenza, M. N.** (2012). Similarities and differences between pathological gambling and substance use disorders: A focus on impulsivity and compulsivity. *Psychopharmacology*, *219*(2), 469–490. https://doi.org/10.1007/s00213-011-2550-7

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

Maia, T. V., & **Frank, M. J.** (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. https://doi.org/10.1038/nn.2723

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

Nebe, S., Kretzschmar, A., Brandt, M. C., & **Tobler, P. N.** (2024). Characterizing Human Habits in the Lab. *Collabra: Psychology*, *10*(1), 92949. https://doi.org/10.1525/collabra.92949

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*. Vienna, Austria: R Foundation for Statistical Computing. 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., Deserno, L., Wilbertz, T., Heinze, H.-J., & **Schlagenhauf, F.** (2016). Risk Factors for Addiction and Their Association with Model-Based Behavioral Control. *Frontiers in Behavioral Neuroscience*, *10*. https://www.frontiersin.org/articles/10.3389/fnbeh.2016.00026

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

Rummery, G. A., & **Niranjan, M.** (1994). *On-line Q-learning using connectionist systems* (Vol. 37, p. 14). Cambridge, UK: University of Cambridge, Department of Engineering.

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

Sebold, M., Nebe, S., Garbusow, M., Guggenmos, M., Schad, D. J., Beck, A., Kuitunen-Paul, S., Sommer, C., Frank, R., Neu, P., Zimmermann, U. S., Rapp, M. A., Smolka, M. N., Huys, Q. J. M., Schlagenhauf, F., & **Heinz, A.** (2017). When Habits Are Dangerous: Alcohol Expectancies and Habitual Decision Making Predict Relapse in Alcohol Dependence. *Biological Psychiatry*, *82*(11), 847–856. https://doi.org/10.1016/j.biopsych.2017.04.019

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., 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

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

Smith, R., Schwartenbeck, P., Stewart, J. L., Kuplicki, R., Ekhtiari, H., & **Paulus, M. P.** (2020). Imprecise action selection in substance use disorder: Evidence for active learning impairments when solving the explore-exploit dilemma. *Drug and Alcohol Dependence*, *215*, 108208. https://doi.org/10.1016/j.drugalcdep.2020.108208

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

Waford, R. N., & **Lewine, R.** (2010). Is perseveration uniquely characteristic of schizophrenia? *Schizophrenia Research*, *118*(1), 128–133. https://doi.org/10.1016/j.schres.2010.01.031

Wagner, B., Mathar, D., & **Peters, J.** (2022). Gambling environment exposure increases temporal discounting but improves model-based control in regular slot-machine gamblers. *Computational Psychiatry*, *6*(1), 142. https://doi.org/10.5334/cpsy.84

Waller, G., Shaw, T., Meyer, C., Haslam, M., Lawson, R., & **Serpell, L.** (2012). Persistence, Perseveration and Perfectionism in the Eating Disorders. *Behavioural and Cognitive Psychotherapy*, *40*(4), 462–473. https://doi.org/10.1017/S135246581200015X

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

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., & **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., 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

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

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. https://doi.org/10.1016/j.bpsgos.2022.03.011

TO CITE THIS ARTICLE:
Brands, A. M., Mathar, D., & Peters, J. (2025). Signatures of Perseveration and Heuristic-Based Directed Exploration in Two-Step Sequential Decision Task Behaviour. *Computational Psychiatry*, 9(1), pp. 39–62. DOI: https://doi.org/10.5334/cpsy.101

Submitted: 21 June 2023
Accepted: 17 December 2024
Published: 11 February 2025

Computational Psychiatry is a peer-reviewed open access journal published by Ubiquity Press.