Nat Hum Behav. Author manuscript; available in PMC 2022 November 19.

Published in final edited form as:
Nat Hum Behav. 2022 August ; 6(8): 1126–1141. doi:10.1038/s41562-022-01346-2.

Explicit knowledge of task structure is a primary determinant of human model-based action

Pedro Castro-Rodrigues[#1,2,3,4], **Thomas Akam**[#2,5], **Ivar Snorasson**[6], **Marta Camacho**[1,a], **Vitor Paixão**[2], **Ana Maia**[1,2,3,7], **J. Bernardo Barahona-Corrêa**[1,2,3], **Peter Dayan**[8,9], **H. Blair Simpson**[6,10], **Rui M. Costa**[2,3,11], **Albino J. Oliveira-Maia**[1,2,3,*]

[1]Champalimaud Clinical Centre, Champalimaud Foundation, Lisbon, Portugal

[2]Champalimaud Research, Champalimaud Foundation, Lisbon, Portugal

[3]NOVA Medical School, NMS, Universidade Nova de Lisboa, Lisbon, Portugal

[4]Centro Hospitalar Psiquiátrico de Lisboa, Lisbon, Portugal

[5]Department of Experimental Psychology, University of Oxford, Oxford, UK

[6]Center for Obsessive-Compulsive & Related Disorders, New York State Psychiatric Institute, New York, USA

[7]Department of Psychiatry and Mental Health, Centro Hospitalar de Lisboa Ocidental, Lisbon, Portugal

[8]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

[9]The University of Tübingen, Tübingen, Germany

[10]Department of Psychiatry, Columbia University, New York, USA

[11]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

[#] These authors contributed equally to this work.

Abstract

[*]Corresponding author: albino.maia@neuro.fchampalimaud.org.
[a]Current address: John Van Geest Center for Brain Repair, University of Cambridge, UK.

Competing interests
JBBC received honoraria from Janssen-Cilag, Ltd, as member of a local Advisory Board. HBS has received research support for an industry-sponsored clinical trial from Biohaven Pharmaceuticals, royalties from UpToDate Inc., and a stipend from the American Medical Association for her role as Associate Editor of JAMA Psychiatry. AJOM was national coordinator for Portugal of a non-interventional study (EDMS-ERI-143085581, 4.0) to characterize a Treatment-Resistant Depression Cohort in Europe, sponsored by Janssen-Cilag, Ltd (2019-2020), is recipient of a grant from Schuhfried GmBH for norming and validation of cognitive tests, and is national coordinator for Portugal of trials of psilocybin therapy for treatment-resistant depression, sponsored by Compass Pathways, Ltd (EudraCT number 2017-003288-36 and 2020-001348-25), and of esketamine for treatment-resistant depression, sponsored by Janssen-Cilag, Ltd (EudraCT NUMBER: 2019-002992-33). The remaining authors declare no competing interests.

Authors contributions
PCR, TA, BBC, HBS, RMC and AJOM conceived and designed the experiments. PCR, IS, MC and AM performed the experiments. PCR and TA analyzed the data. TA, VP, PD, RMC and AJOM contributed materials/analysis tools. PCR, TA and AJOM wrote the paper.

Europe PMC Funders Author Manuscripts

Explicit information obtained through instruction profoundly shapes human choice behaviour. However, this has been studied in computationally simple tasks, and it is unknown how model-based and model-free systems, respectively generating goal-directed and habitual actions, are affected by the absence or presence of instructions. We assessed behaviour in a variant of a computationally more complex decision-making task, before and after providing information about task structure, both in healthy volunteers and individuals suffering from obsessive-compulsive (OCD) or other disorders. Initial behaviour was model-free, with rewards directly reinforcing preceding actions. Model-based control, employing predictions of states resulting from each action, emerged with experience in a minority of participants, and less in OCD. Providing task structure information strongly increased model-based control, similarly across all groups. Thus, in humans, explicit task structural knowledge is a primary determinant of model-based reinforcement learning, and is most readily acquired from instruction rather than experience.

Introduction

The brain uses multiple systems to choose which actions to perform[1–6]. One widely held distinction is made between goal-directed actions, guided by predictions of their specific outcomes, and habitual actions, performed according to preferences acquired through prior reinforcement[1,7–9]. This cognitive and behavioural classification is thought to correspond, at least in part, to a computational distinction between two different types of reinforcement learning (RL), termed model-based and model-free[4,5,10,11]. A model-based RL algorithm learns to predict the specific consequences of actions, and computes their values, i.e. long run utilities, by simulating likely future behavioural trajectories. This allows for statistically efficient use of experience, and thus flexibility, at the cost of the computational demands of planning[4,10,12]. Model-free RL, by contrast, learns estimates of the value of states or actions directly from experience, and updates these estimates using reward prediction errors. This allows for rapid action selection at low computational cost, but uses information less efficiently, resulting in slower adaption to changes in the environment[4,10]. It is thought that the brain takes advantage of the complementary strengths of both prospective (model-based) and retrospective (model-free) approaches to decision-making, through mechanisms that estimate whether the payoff for more accurate prediction is worth the computational costs of planning[4,13,14].

Sequential, or multi-step, decision tasks have emerged as a powerful approach to study model-based and model-free RL in humans[11,13,15,16]. In such tasks, participants move through a sequence of states to obtain rewards, typically with non-stationary reward and/or action-state transition probabilities, forcing continuous learning. The contributions of model-based and model-free RL can be determined by examining how participants update their choices in light of recent experience. To date, the most commonly used task is the ‘two-step’ task, employing a choice between two ‘first-step’ stimuli, which leads probabilistically to one of two ‘second-step’ states, where rewards may be obtained[11]. Each first-step stimulus commonly leads to one of the second-step states but, on a minority of trials, leads to the state commonly reached from the other stimulus. Model-based and model-free RL are identified according to how the trial outcome (rewarded or not) and state transition (common or rare) interact to affect the subsequent choice. Under model-free control, the

agent will tend to repeat first-step choices that are followed by reward, irrespective of the state transition. By contrast, under model-based control, the agent will tend to switch first-step choice when a rare transition leads to reward, given that the reward increases the value of the state commonly reached from the not-chosen first step option. The two-step task has been used to study neural correlates of model-based and model-free control in healthy participants[11,17–27], and to investigate decision making in clinical populations[28–32].

Human participants typically receive extensive instruction about task structure prior to performing the two-step task[11,33]. However, though there is extensive literature showing that instruction profoundly shapes human behaviour in operant[34–36] and fear[37,38] conditioning, as well as value-based decision making[39–42], little is known about how instruction affects behaviour in multi-step tasks that dissociate model-based and model-free control. To our knowledge, a single study in healthy humans has partially addressed this question, showing that making instructions more comprehensive and easier to understand, increases the influence of model-based relative to model-free RL[33]. This result, in combination with other analyses, led the authors to propose that humans are primarily model-based learners on this task, suggesting that apparent model-free behaviour, including in some clinical populations, may result from incorrect task models, rather than a true model-free strategy.

However, few studies have explored behaviour on multi-step tasks in the absence of information about task structure, in either healthy or clinical populations. Thus, it remains unclear how model-based and model-free RL contribute to action selection in situations where participants must learn task structure directly and exclusively from experience, and how providing explicit information about task structure affects each system. To address these questions, we created a simplified version of the two-step task, requiring minimal prior instruction. We initially administered it with no information given about the task state space, transition structure, or reward probabilities. Then, the task was repeated following debriefing about these elements of the task's structure. Behaviour was tested in healthy volunteers, as well as in a sample of individuals with obsessive-compulsive disorder (OCD), previously reported to have deficits in the degree to which model-based RL is employed[29,31]. To control for the effects of psychotropic medication and of unspecific mood and anxiety symptoms, data was also collected in a comparison sample of individuals with mood and anxiety disorders. Initial behaviour, prior to instructions, was model-free across all groups, with model-based control emerging with experience in only a minority of participants, and to a lesser extent in OCD. However, once task structure information was provided, model-based control increased to a very similar, and significant extent in healthy volunteers and individuals with OCD or other disorders. These findings demonstrate that explicit task structural knowledge is a primary determinant of human use of model-based RL, and is most readily acquired from instruction rather than experience.

Results

We developed a simplified two-step task requiring minimal prior instruction. Specifically, we simplified the visual representation of task states on the screen, the task structure (allowing only a single action rather than a choice in each second-step state), and the reward probability distribution (using blocks, instead of slowly fluctuating Gaussian random

walks, to increase the contrast between good and bad options[43,44]; Figure 1). Two-hundred and four individuals were recruited in Lisbon and New York to perform this task: 109 were healthy volunteers, 46 were diagnosed with OCD and 49 with other mood and anxiety disorders. Sociodemographic and psychometric data from all participants is shown in table 1. While statistically significant differences between groups were not found for age (F_2=2.4, P=0.1, η^2=0.02, one-way ANOVA) nor gender (χ^2=5.1, P=0.3, Cramer's V=0.1, Pearson's chi-squared), they were present for years of education (F_2=5.7, P<0.01, η^2=0.06, one-way ANOVA), which were slightly, but significantly, higher in healthy volunteers than the mood and anxiety group (difference between means=1.7, 95% CI [0.42, 2.91], P=0.01, Tukey's HSD). As expected, both clinical groups had significantly higher anxiety and depression scores than healthy volunteers (F_2>55, P<0.001, 0.36<η^2<0.62, across all one-way ANOVA's for depression and anxiety scores), while participants with OCD had higher obsessive-compulsive scores than participants in either of the two other groups (F>200, P<0.001, 0.76<η^2<0.81, across all one-way ANOVA's for the Yale-Brown Obsessive-Compulsive Scale total and sub-scores). Regarding medication, we classified it in classes and we did not find statistically significant differences between clinical groups in the use of any class of medication (χ^2<1.6, P>0.2, across all Chi-squared tests).

While developing the task among healthy volunteers in Lisbon, participants were randomized between two different versions, one with fixed transition probabilities linking the first-step actions and second-step states (Fixed version; n=40), and one where the transition probabilities underwent periodic reversals (Changing version; n=42). The Changing version proved too complex for most participants, particularly as shown by the lack of effects of debriefing on the development of model-based RL, although a small subset was able to learn the task structure (see "Changing transition probabilities inhibits model-based control" in Supplementary Information for details), so we subsequently focused on the Fixed task, which is used for all data and figures in the main text. All healthy volunteers recruited in New York (n=27), and all clinical participants at both sites (n=95) completed this version. All participants in both versions performed 4 sessions of 300 trials each in a single day. A subset of healthy controls, and all clinical participants, were debriefed between sessions 3 and 4, with the task structure explained to them. We assessed the effect of uninstructed experience by comparing behaviour between sessions 1 and 3, and the effect of explicit knowledge by comparing behaviour between sessions 3 and 4.

Initial behaviour is under model-free control

As participants were not told how their actions (arrow key presses) affected the stimuli shown on the screen, they had to learn both the correspondence between arrow keys and stimuli, and that stimuli could only be selected when highlighted. In the 67 healthy volunteers performing the Fixed version of the task, the number of invalid key presses per trial (i.e. presses to keys whose corresponding stimulus was not highlighted) decreased over the first 50-100 trials, before stabilising at a low level in all but a minority of participants (Supplementary Figure 1). During session 1 there was no statistically significant difference in the average rate of invalid key presses at the second-step following common (median=0.027/trial) vs rare (median=0.023/trial) transitions (P=0.2, Sign test, Supplementary Figure 1).

Europe PMC Funders Author Manuscripts

To assess how trial events affected subject's choices, we analysed the probability of repeating first-step choices (termed 'stay probabilities') as a function of the previous state transition (common or rare), trial outcome (rewarded or not), and their interaction[11]. During session 1, stay probability was strongly influenced by trial outcome (coefficient=1.17, 95% CI [0.92,1.44], P<0.001, bootstrap test). There was no statistically significant evidence for an influence of state transition (coefficient=0.07, 95% CI [-0.07,0.21], P=0.3) or the transition-outcome interaction (coefficient=0.1, 95% CI [-0.02,0.23], P=0.1); Figure 2 a, d) on stay probability. This pattern is consistent with a simple model-free strategy, in which the outcome received at the end of the trial directly reinforces the choice made at the first-step[11]. This direct reinforcing effect of reward was evident from the very start of the first session (Figure 2b), rather than emerging with task experience.

Although we did not find evidence for state transitions influencing subsequent first-step choices in session 1, key-press reaction times at the second-step were faster following common than rare transitions (399.1 ± 16.9ms and 514.4 ± 20.5ms respectively; t_{66}=7.81, P<0.0001, d=0.75, paired t-test; Figure 2e). This dissociation between choice and implicit measures of task-structure learning suggests that motor systems learned to predict and prepare upcoming actions before decision making systems were using a predictive model to evaluate choices.

Modest increase in model-based control with experience

To assess how task experience affected behavioural strategy we compared behaviour in sessions 1 and 3. Stay probability at session 3 was more strongly influenced by both state transition (null 95% CI [-0.18,0.18], coefficient change=0.27, P=0.003, permutation test) and the transition-outcome interaction (null 95% CI [-0.25,0.24], coefficient change=0.39, P<0.001), while evidence for a change in the influence of trial outcome was not statistically significant (null 95% CI [-0.31,0.32], coefficient change=0.31, P=0.06; Figure 2 c, d). This pattern is consistent with increased influence of model-based control, as model-based agents know that outcomes following rare transitions primarily influence the value of the first-step option that was not chosen[11,43], leading to loading on the transition-outcome interaction predictor. Importantly, loading on the transition–outcome interaction parameter across sessions 1 to 3 was positively correlated with the number of rewards obtained by each subject (r[65]=0.41, P<0.001; 95% CI [0.19, 0.59], Pearson's correlation), suggesting that participants who learned a model of the task used information more efficiently and thus obtained rewards at a higher rate. Furthermore, we have previously shown that, when transition probability estimates are updated based on experienced state transitions, as is the case here, model-based agents tend to repeat the same choice after common transitions, producing a positive coefficient for state transition as a predictor of stay probability[43]. Increased loading on the state transition predictor thus provides added support for development of model-based control with task experience between sessions 1 and 3.

To further explore model-free and model-based control prior to receiving instructions, we fit RL models to the data. Model-comparison combining data from sessions 1-3 indicated that a mixture model including model-free and model-based components fit data better than a purely model-free or a purely model-based model, as reflected by lower Bayesian

Information Criteria (BIC) scores for the mixture model (Supplementary Figure 2a, left panel). Models that included a "bias" parameter, capturing bias towards the upper or lower first-step choice, and a "perseveration" parameter, capturing a tendency to repeat the previous choice, fit the data better than a model not including these parameters (Supplementary Figure 2a, right panel). As it has been suggested that apparently model-free behaviour could in fact reflect a model-based strategy with an incorrect model of the task structure[33], we considered 3 additional model-based agents with incorrect beliefs, but found these fit the data from both session 1 and 3 worse than any of the traditional models (Supplementary Figure 2b). We also simulated behaviour from the best fitting RL model and verified that it produced stay probability plots qualitatively similar to the experimental data (Supplementary Figure 5).

To assess uninstructed learning effects using RL models, we compared parameter values of the fitted models for sessions 1 and 3. The value learning rate increased significantly between session 1 and 3 (null 95% CI [-0.17,0.17], parameter change=0.18, P=0.03), permutation test), but there was no statistically significant evidence for changes other parameters (P>0.06), including the strength of model-based influence on choices (null 95% CI [-0.30,0.30], parameter change=0.07, P=0.65; Figure 2f). The discrepancy with increased loading on the 'transition x outcome' predictor in the stay-probability analysis may reflect lower statistical power to detect subtle strategy changes in the strongly non-linear and more flexibly parameterised RL model. It likely also reflects the fact that only a minority of participants learned to use model-based RL, with per-subject model comparison between the mixture RL model and a simpler model-free RL model indicating that only 15% of participants (10/67) used model-based RL at session 3 (likelihood ratio test, threshold P=0.05).

Key-press reaction times at the second-step became faster overall between session 1 and 3 (main effect of session $F_{1,66}$=21.1, P<0.0001, η_p^2=0.24), but this was more pronounced following common than rare transitions (session-transition interaction $F_{1,66}$=21.1, P=0.008, η_p^2=0.1, repeated measures ANOVA; Figure 2e). Additionally, by session 3 the rate of invalid key presses was significantly higher following rare (median=0.037/trial) than common (median=0.017/trial) transitions (P=0.004, Sign test, Supplementary Figure 1). Therefore, both choice-based and implicit measures showed evidence of learning about the transition structure between session 1 and 3. The strength of model-based influence on choice was significantly correlated across participants with the rare-common reaction time difference at both session 1 (r[65]=0.57; P<0.001, 95% CI [0.37,0.71]; Pearson's correlation, Supplementary Figure 3) and session 3 (r[65]=0.69; P<0.001, 95% CI [0.54,0.80]), suggesting interaction between learning at motor and cognitive-levels, though this appeared to be driven by the minority of participants whose choices were more model-based.

A possible reason why model-free control might predominate is that participants could perform the task as fast as they wished and, thus, might have been optimising speed over accuracy. To address this possibility, we tested an additional group of 20 healthy volunteers (mean age = 29.6 years old [SD=9]; gender = 25% males; mean education = 15.3 years [SD = 2.8]) on a slow-paced version of the task, in which a 1 second delay occurred between

circles lighting up and being active for selection, cued by a change in colour from pale to bright yellow, in addition to a 1 second intertrial interval (ITI). Participants completed three sessions, each of 150 trials followed by receiving explicit information about task structure, and a further session of 150 trials afterwards. As in the self-paced task, initial behaviour was consistent with model-free control, with a main effect of trial outcome on stay probability in session 1 (coefficient=0.79, 95% CI [0.44,1.14], $P<0.001$, bootstrap test). There was no statistically significant evidence for an effect of transition (coefficient=0.16, 95% CI [-0.10,0.41], $P=0.2$) or the transition-outcome interaction (coefficient=-0.01, 95% CI [-0.24,0.25], $P=0.9$) on stay probability (Supplementary Figure 4a,b). Also, similarly to the self-paced task, the effect of transition-outcome interaction on stay probability increased between session 1 and 3 (null 95% CI [-0.35,0.34], coefficient change=0.45, $P=0.005$, permutation test), as assessed by the logistic regression (Supplementary Figure 4b), consistent with increased use of model-based control with experience. However, at session 3 the influence of model-free control was still substantially larger than that of model-based, as assessed by RL model-fitting (Supplementary Figure 4d), and a likelihood ratio test on session 3 data supported a mixed model-based plus model-free strategy over a simpler model-free only strategy in only 3 among the 20 participants.

Overall, these data show that while signatures of model-based RL increased modestly with uninstructed experience, model-free RL predominated during uninstructed behaviour in this unfamiliar domain, and remained the strongest influence on choices for most participants over uninstructed trials.

Impaired use of model-based control with experience in OCD

Ninety-five individuals with either OCD or mood and anxiety disorders (Table 1) also completed the Fixed version of the simplified two-step task. In the stay probability analysis, when comparing session 1 with session 3, we did not find statistically significant evidence for an increased influence of transition (null 95% CI [-0.23,0.23], coefficient change=0.2, $P=0.09$, permutation test) or transition-outcome interaction (null 95% CI [-0.23,0.22], coefficient change=-0.04, $P=0.7$) in the OCD group (n=46; Figure 3a, b). Instead, there was an increased influence of trial outcome over uninstructed learning (null 95% CI [-0.35,0.36], coefficient change=0.58, $P<0.001$), that may reflect enhanced model-free control with experience. However, in direct comparisons with healthy volunteers (n=67), session by group interaction was significant only for the transition-outcome interaction parameter (null 95% CI [-0.35,0.36], group difference in coefficient change=-0.43, $P=0.015$), but not for the transition (null 95% CI [-0.30,0.30], group difference=-0.07, $P=0.65$) or outcome parameters (null 95% CI [-0.46,0.48], group difference=0.27, $P=0.25$).

As in healthy participants, second-step reaction times were faster following common than rare transitions (main effect of transition, $F_{1,45}=51.3$, $P<0.0001$, $\eta_p^2=0.53$, repeated measures ANOVA; Figure 3c), and also faster in session 3 than session 1 (main effect of session, $F_{1,45}=10$, $P=0.003$, $\eta_p^2=0.18$). However, the session by transition interaction did not reach significance ($F_{1,45}=1.95$, $P=0.16$, $\eta_p^2=0.04$). Directly comparing OCD and healthy volunteers, while individuals with OCD had slower reaction times overall (main effect of group, $F_{1,111}=8.65$, P=0.004, $\eta_p^2=0.07$, mixed ANOVA), interactions with group were

not significant for session, transition, nor session by transition interaction (all $F_{1,111}<0.67$, $P>0.4$, $\eta_p^2<0.006$). Finally, consistent with the stay probability analysis, RL mixture model fits to sessions 1 and 3 (Figure 3c) showed an increase in the influence of model-free action values on choice over learning (null 95% CI [-1.36,1.32], parameter change=1.71, $P=0.012$, permutation test), although session-by-group interaction with healthy volunteers did not reach significance (null 95% CI [-1.68,1.68], group difference in parameter change=1.52, $P=0.07$, permutation test, Supplementary Table 1).

To investigate potential contributions of medication or of unspecific mood and anxiety symptoms for the findings in the OCD group, equivalent experiments and comparisons were performed in a sample of individuals with other mood and anxiety disorders (n=49). Here, in stay probability analysis (Figure 3e, f) we found an increased influence of trial outcome (null 95% CI [-0.33,0.35], coefficient change=0.63, P<0.001, permutation test) and transition (null 95% CI [-0.25,0.25], coefficient change=0.33, P=0.011) with task experience, but we did not find statistically significant evidence for a change in the influence of the transition-outcome interaction predictor on stay probability (null 95% CI [-0.26,0.27], coefficient change=0.20, P=0.15). Second-step reaction times were faster following common than rare transitions (main effect of transition, $F_{1,48}=34.2$, $P<0.0001$, $\eta_p^2=0.42$, repeated measures ANOVA, Figure 3g), and faster in session 3 than session 1 (main effect of session, $F_{1,48}=30.5$, $P<0.0001$, $\eta_p^2=0.39$), but the session by transition interaction was not statistically significant ($F_{1,48}=0.86$, $P=0.36$, $\eta_p^2=0.02$). Compared with the healthy volunteers, this group had slower second-step reaction times overall (main effect of group, $F_{1,114}=6.97$, $P=0.009$, $\eta_p^2=0.06$, mixed ANOVA), and a stronger influence of both the transition type (group-transition interaction $F_{1,114}=4.26$, $P=0.041$, $\eta_p^2=0.04$) and session number (group-session interaction $F_{1,114}=6.74$, $P=0.011$, $\eta_p^2=0.06$) on reaction time. Finally, RL model fits showed only an increased value learning rate (null 95% CI [-0.21,0.21], coefficient change=0.26, P=0.011, permutation test). We did not find statistically significant evidence for changes in the influence of model-free or model-based action values on choice over learning (model-free: null 95% CI [-0.80,0.80], coefficient change=0.49, P=0.24, model-based: null 95% CI [-0.43,0.42], coefficient change=0.10, P=0.64 permutation test) (Figure 3h). Importantly, there were no statistically significant session by group interactions between these patients and healthy volunteers for the stay probability analysis or RL model fits (P>0.11, Supplementary Table 2, permutation test). Overall, these data suggest a different pattern of learning from experience in individuals with OCD, with a failure to learn the task-transition structure and exhibit model-based RL.

Explicit knowledge increases model-based control

We next assessed how providing explicit information about the task structure changed behaviour, by comparing behaviour in sessions 3 and 4 in a group that received debriefing about task structure after session 3, and in another group that was not provided such information. To avoid ceiling effects in participants who already acquired a model of the task, these analyses only included the 57 healthy volunteers for whom a likelihood ratio test indicated model-based RL was not being used significantly in session 3, as described above. Among these participants, in session 4, more than 50% of those that were debriefed were identified by the likelihood-ratio test as using model-based RL (21/41), while in

the absence of debriefing, only one subject became model-based (1/16; z=3.13, *P*=0.002, z-test for difference in proportions; Figure 4a, f). Consistently, debriefing strongly affected how events on each trial influenced the subsequent choice (Figure 4b, c, g, h), with increased influence of state transition (null 95% CI [-0.42,0.42], coefficient change=0.75, *P*<0.001; null 95% CI [-0.61,0.60], debriefing vs no-debriefing group difference in coefficient change=0.66, *P*=0.03; permutation tests) and transition-outcome interaction (null 95% CI [-0.51,0.50], coefficient change=1.07, *P*<0.001; null 95% CI [-0.73,0.67], group difference=0.93, *P*=0.002) on stay probability. Similar effects of debriefing on the transition-outcome interaction were found among the 17 healthy volunteers performing the slow-paced version of the task (Supplementary Figure 4e-g).

RL mixture model fits of pre and post debriefing data (Figure 4e, j) confirmed that the influence of model-based action values on choice was increased by debriefing (null 95% CI [-0.70,0.70], parameter change=1.17, *P*<0.001; null 95% CI [-1.07,0.92], group difference=1.19, *P*=0.006). Furthermore, the influence of model-free action values on choice reduced after debriefing (null 95% CI [-0.79,0.79], parameter change=-1.04, *P*=0.006), while value learning rates increased (null 95% CI [-0.18,0.18], parameter change=0.29, *P*<0.001), though the session by group interactions were not statistically significant (respectively: null 95% CI [-1.73, 2.15], group difference=-0.36, *P*=0.7; and null 95% CI [-0.30, 0.32], group difference=0.07, *P*=0.75). In addition to modifying choice behaviour, debriefing increased differences in second-step key-press reaction times between common and rare transition trials (debriefing group session-transition interaction $F_{1,40}$=59.6, P<0.0001, η_p^2=0.59, repeated measures ANOVA; session-transition-group interaction $F_{1,55}$=19.3, P<0.0001, η_p^2=0.26, mixed ANOVA comparing debriefing and non-debriefing groups, Figure 4d), further supporting that the influence of state transition on RT in this task comprises both a motor component, which is independent of the use of model-based RL, and a cognitive component which manifests when participants are using model-based RL. Statistically significant evidence for differences in comparisons between sessions 3 and 4 were not found in the no debriefing group (Figure 4f-j). Similar effects of debriefing on model-based action values, value learning rates and second-step key-press reaction times were found among healthy volunteers performing the slow-paced version of the task, albeit the sample size not being sufficient to produce significant differences in some comparisons (Supplementary Figure 4h,i).

Finally, among participants recruited in Lisbon, where neuropsychological data was available, we tested for correlations between test scores, namely from the Corsi block tapping test (assessing visuospatial working memory) and a Go/No-Go task (number of No-Go errors and reaction-time, assessing impulsivity), with several behavioural measures, specifically the outcome and transition-outcome interaction logistic regression predictor loadings, as well as the RL model parameters controlling the influence of model-free and model-based values on choice. Significant correlations were not found, neither among all healthy volunteers using data from session 3 (-0.27<r[38]<0.31, 0.054<*P*<0.8, %95 CI [-0.54<lower bound<-0.01, 0.05<upper bound<0.57]; Pearson's correlation), nor among the debriefing group using data from session 4 (-0.45<r[15]<0.38, 0.07<*P*<0.8, %95 CI [-0.76<lower bound<-0.16, 0.08< upper bound<0.75]).

Explicit knowledge affects value updates and perseveration

Unexpectedly, the RL model eligibility trace parameter also decreased after debriefing (null 95% CI [-0.16,0.17], parameter change=-0.23, P=0.006; null 95% CI [-0.31,0.28], group difference=-0.36, P=0.024; permutation tests; Fig. 4e). A similar effect of debriefing on the eligibility trace parameter was found among healthy volunteers performing the slow-paced version of the task (null 95% CI [-0.26,0.27], parameter change=-0.27, P=0.048; Supplementary Figure 4i). This parameter controls the relative influence of the second-step state's value and the trial outcome on updates to model-free first-step action values. Debriefing increased the influence of the second-step state value and decreased that of the trial outcome. As there is no obvious reason why providing task structure information should change model-free eligibility traces, we hypothesized that this effect is in fact mediated by the influence of task structural knowledge on representation of the task state space. By telling participants that the reward probabilities depend on the second-step state reached, these states are likely made more distinct and salient in their internal representation of the task, and hence better able to accrue value, which can then drive model-free updates of first step action values. Consistent with this interpretation, participants who, following debriefing, had large increases in the strength of model-based control, indicating that they had correctly understood the task structure, also had a larger decrease in the eligibility trace parameter (r[39]=-0.34, P=0.03; 95% CI [-0.59, -0.04]; Pearson's correlation; Supplementary Figure 6).

Debriefing also increased how often participants repeated choices independent of subsequent trial events, as reflected by a significant increase in the 'perseveration' parameter of the RL model (null 95% CI [-0.75,0.76], parameter change=1.63, P<0.001; null 95% CI [-1.25,1.07], group difference in parameter change=1.76, P<0.001; permutation tests; Fig. 4e). This may result from information that reward probabilities on the left and right reversed only occasionally and are thus stable for extended periods of time. In this case, one would expect a reduction in perseveration across the course of each block, from shortly after a reversal, when reward probabilities are stable, to late in the block, when the next reversal is anticipated. Consistent with this hypothesis, we found that participants with larger post-debriefing increases in overall perseveration also had larger declines in perseveration within post-debriefing non-neutral blocks, from trials 10-20 (early) to 30-40 (late; r[39]=-0.35, P=0.02; 95% CI [-0.59, -0.05]; Pearson's correlation; Supplementary Figure 6).

To verify that changes in other model parameters (e.g. MF and MB weights) had not artifactually caused these effects by preventing us from accurately estimating parameter values, we assessed the accuracy of parameter recovery from simulated data (Supplementary Figure 7). Overall, the accuracy of parameter recovery was very good, with a slightly reduced accuracy for the transition probability learning rate (parameter αT) in sessions 1 and 3, where the influence of model-based RL is small. Furthermore, we tested for differences in learning or debriefing effects between the Lisbon and New York debriefing groups, and did not find any significant differences (Supplementary Table 3).

Europe PMC Funders Author Manuscripts

Explicit knowledge increases model-based control in OCD

In the 41 of 46 individuals with OCD that were model-free at session 3, 37% (15 participants) started using model-based RL after debriefing (Figure 5a, likelihood ratio test with threshold P=0.05). Consistent with this, the stay probability analysis showed increased loading on both the transition (null 95% CI [-0.35,0.37], coefficient change=0.56, $P<0.001$, permutation tests) and transition-outcome interaction (null 95% CI [-0.50,0.51], parameter change=0.88, $P<0.001$) predictors, similarly to that observed in healthy controls (Figure 5b, c). Increased use of model-based RL after debriefing was confirmed by model fitting (Figure 5e), which showed increased influence of model-based action values on choice (null 95% CI [-0.63,0.64], parameter change=1.22, $P<0.001$), and a trend towards reduced influence of model-free action values (null 95% CI [-1.16,1.18], parameter change=-1.08, $P<0.07$). As in healthy volunteers, debriefing in participants with OCD increased differences in second step reaction times between common and rare transition trials (session-transition interaction $F_{1,40}=30.8$, P<0.0001, $\eta_p^2=0.43$, repeated measures ANOVA; Figure 5d), while overall reaction times remained slower than in healthy volunteers (main effect of group, $F_{1,80}=7.31$, P=0.008, $\eta_p^2=0.08$, mixed ANOVA). Again, similarly to healthy volunteers, debriefing reduced the value of the eligibility trace parameter (null 95% CI [-0.19,0.19], parameter change=-0.24, $P=0.017$), and this decrease correlated with increased use of model-based RL (r[39]=-0.56, P=0.0001; 95% CI [-0.74, -0.30]; Pearson's correlation; Supplementary Figure 6). Though debriefing increased choice perseveration in OCD participants (null 95% CI [-0.67,0.66], parameter change=0.77, $P=0.023$), the effect was significantly smaller than in healthy volunteers (null 95% CI [-0.84,0.8], group difference in parameter change=-0.86, $P=0.042$). This was the only significant interaction between debriefing and OCD diagnosis in direct comparisons with data from healthy volunteers for stay probability analysis and RL model fits (Supplementary Table 1). In individuals with OCD we did not find statistically significant correlation between the increase in perseveration following debriefing and changes in perseveration from early to late in blocks after debriefing (r[39]=-0.09, $P=0.5$; 95% CI [-0.39, 0.22]; Pearson's correlation; Supplementary Figure 6).

In the group with mood and anxiety disorders, among 37 participants that were model-free at session 3, 68% (25 participants) started using model-based RL after debriefing (Figure 5f, likelihood ratio test with threshold P=0.05).debriefing increased the influence of transition (null 95% CI [-0.40,0.41], coefficient change=0.64, $P<0.001$, permutation tests) and transition-outcome interaction (null 95% CI [-0.57,0.59], coefficient change=1.20, $P<0.001$) predictors on stay probability (Figure 5f, g), similarly to healthy volunteers. The RL model fit confirmed that debriefing increased the influence of model-based action values on choice (null 95% CI [-0.81,0.81], parameter change=1.63, $P<0.001$), and reduced influence of model-free action values (null 95% CI [-0.69,0.71], parameter change=-0.82, $P=0.019$; Figure 5h). As in healthy volunteers, debriefing increased the difference in second-step reaction times between common and rare transition trials ($F_{1,36}=26.2$, P<0.0001, $\eta_p^2=0.42$, repeated measures ANOVA), and there was no statistically significant difference in reaction time effects between this group and healthy controls ($F_{1,76}<2.9$, P>0.10, $\eta_p^2<0.04$, mixed ANOVA). Debriefing effects observed in healthy volunteers for the value learning rate (null 95% CI [-0.17,0.18], parameter change=0.21, $P=0.015$) and eligibility trace parameters(null 95% CI [-0.15,0.15], parameter change=-0.15, $P=0.04$) were also replicated here, with

decrease in the latter again correlating with increased use of model-based RL (r[35]=-0.45, P=0.005; 95% CI [-0.68, -0.15]; Pearson's correlation; Supplementary Figure 6). However, unlike in the healthy volunteers, we did not find a statistically significant effect of debriefing on choice perseveration (null 95% CI [-0.46,0.47], parameter change=0.23, P=0.35; null 95% CI [-0.82,0.84], group difference=-1.4, P=0.001), and no significant correlation across participants between the debriefing effect on perseveration and post debriefing change in perseveration from early to late in blocks (r[35]=-0.16, p=0.36, 95% CI [-0.46, 0.17]; Pearson's correlation). This was the only significant session by group interaction for stay probability analysis and RL model fits (Supplementary Table 2). Overall, there was no significant evidence that either clinical group was less able to use information about task structure to employ a model-based strategy, but both clinical groups showed a reduced influence of this information on choice perseveration.

To further explore potential effects of medication, we also tested for differences in RL strategies between the clinical groups recruited in Lisbon, the majority of whom were receiving pharmacological treatment (13/16 in OCD group, 14/16 in mood and anxiety disorders group), and the clinical groups recruited in New York, who were tested in the absence of such treatment. We found that debriefing reduced the strength of model-free RL in treated but not untreated individuals with OCD (null 95% CI [-2.86,2.5], group difference in parameter change=-3.12, P=0.037; Supplementary Table 4, Supplementary Figure 8). Among individuals with other mood and anxiety disorders, significant differences were not found between Lisbon and New York samples (Supplementary Table 5).

Finally, among clinical groups recruited in Lisbon, where neuropsychological data was available, we tested correlations between test scores from the Corsi block tapping test or a Go/No-Go task, and outcome or transition-outcome interaction logistic regression predictor loadings, as well as model-free or model-based the RL model parameters, as described above for healthy volunteers. In the group with mood and anxiety disorders there were significant positive correlations between reaction time in the Go/No-Go task and several measures of model-based control, namely the transition-outcome interaction predictors from sessions 3 (r[14]=0.69; P=0.007; 95% CI [0.3, 0.88]; Pearson's correlation), and 4 (r[14]=0.54; P=0.048, 95% CI [0.06, 0.82]), and the fitted model-based strength parameter value from session 4 (r[14]=0.58; P=0.03, 95% CI [0.12, 0.84]). Other correlations were not statistically significant (-0.26<r[14]<0.31, 0.059<P<0.99, %95 CI [-0.69<lower bound<-0.01, 0.31<upper bound<0.82]; Pearson's correlation),.

Discussion

We developed a simplified two-step task to examine how model-based and model-free RL contribute to behaviour in healthy and clinical populations, when task structure must be learned directly from experience. This allowed for subsequent testing of modifications of behavioural strategies once information about task structure was provided. In healthy volunteers, uninstructed behaviour was initially model-free, with strong direct reinforcement of choices by rewards from the start of the first session, but no evidence of participants using knowledge of task structure early on. In fact, even with extensive experience, signatures of model-based control increased only modestly at the population level, and unevenly across

participants. This is striking given the relative simplicity of the task and suggests that humans are surprisingly poor at learning causal models from experience when they lack prior expectations about task structure. Very similar effects were observed in another group of healthy volunteers tested in a slow-paced version of the task, suggesting that the initial predominance of model-free control was not simply an effect of optimising speed over accuracy. When learning from experience, individuals with OCD were impaired in their acquisition of model-based control, as compared with healthy volunteers. Providing explicit information about task structure strongly increased the use of model-based control, across all tested populations, including individuals with OCD, with additional unexpected effects on model-free action value updates and choice perseveration. The absence of a model-based control deficit in OCD following debriefing is surprising, given the compelling evidence for such deficits in the original two-step task, and raises the question of which task properties determine model-based control deficits in OCD.

The increasing influence of model-based RL with experience contrasts with habit formation in rodent instrumental conditioning, where actions are initially goal-directed but become habitual with extended experience[45], a process thought to involve a transition from model-based to model-free control[4]. This transition, and arbitration between model-based and model-free control more generally, has been proposed to occur through meta-cognitive mechanisms which assess whether the benefits of improved prediction accuracy are worth the costs of model-based evaluation[4,13,14]. The different trajectory in the current task likely results from a more complex state space that increases model uncertainty in early learning and makes model-based learning more demanding, and from ongoing changes in reward probability that prevent the model-free system from converging to accurate value estimates in late learning[4]. In fact, it has been recently suggested that performance during initial stages of action selection tasks may be primarily based on trial-and-error exploration, with progression towards model-based RL occurring in intermediate stages, as participants acquire a model of the environment[46].

Our finding that model-free RL dominates uninstructed behaviour on a two-step task contrasts with recent arguments from Silva & Hare[33] suggesting humans are primarily model-based learners on two-step tasks, and that apparent model-free behaviour is in fact model-based control using muddled or incorrect task models. We cannot rule out the possibility that some apparently model-free behaviour at later uninstructed sessions in our task was in fact model-based control with an incorrect model, though model comparison did not favour any of the incorrect-model strategies proposed. However, we do not think this is a plausible overall explanation for the observed predominance of model-free behaviour prior to instructions. Firstly, because stay probabilities at session one showed a strong main effect of outcome but no statistically significant evidence for a transition-outcome interaction, i.e. the canonical picture of a model-free agent. This is not consistent with Silva and Hare's simulations of agents with muddled models, which show a strong effect of transition-outcome interaction[33]. Secondly, our participants showed a direct reinforcing effect of reward on first-step choice from their very first interactions with the task. It does not appear likely that participants almost instantly acquire muddled models of the task which happen to produce the exact effect predicted by model-free reinforcement. Rather, we propose that, consistent with findings as early as Thorndike's law of effect[47], rewards

in our task had a direct reinforcing effect on actions performed shortly prior to their being obtained. While these findings provide evidence that participants use model-free control in unfamiliar domains, this speaks only indirectly to the question of whether model-free RL or muddled models underlie apparent model-free behaviour in the original two-step task.

Providing explicit information about task structure strongly boosted the influence of model-based RL. This complements Silva and Hare's findings that model-based control is increased by making instructions more complete and embedding them in a narrative to make them easier to remember and understand. Such instruction effects are consistent with meta-cognitive cost-benefit decision making, since an accurate model of the task structure will boost the estimated accuracy of model-based predictions and hence the expected payoffs from model-based control. Our findings also build on extensive literature examining how instruction and experience interact to determine human behaviour, in tasks that do not discriminate model-free and model-based control. Early work examining instruction effects on operant conditioning found that after explicit information about the schedule of reinforcement, responses match the contingencies explained to participants (e.g. fixed interval, variable interval or fixed ratio), even when these differ substantially from the actual contingencies[34–36]. In common with our study, these results emphasize that humans learn about task structure much more readily from explicit information than via trial-and-error learning. More recent work has focused on the effect of advice, i.e. informing participants that one option is particularly good or bad, on reward guided decision making in probabilistic settings[39,40]. Such advice impacts not only initial estimates of how good or bad different options are, but also modifies subsequent learning, by up-weighting and down-weighting outcomes. Whether such bias effects extend to learning about task structure, in addition to simple reward learning, is an open question for further work. Functional neuroimaging has also shown that instructions change responses to outcomes in the striatum, ventromedial prefrontal cortex and orbitofrontal cortex, potentially mediated by representations of instructed knowledge in the dorsolateral prefrontal cortex[38,41,48]. Our task provides a potential tool for extending such mechanistic investigation of instruction effects into the domain of task structure learning and model-based RL.

Our findings may have translational relevance for OCD. Prior studies have shown that individuals with OCD, as well as healthy volunteers with self-reported OCD-like symptoms, have deficits in model-based control in the original two-step task[29,31]. There is also data showing that these findings reflect a transdiagnostic compulsivity dimension, rather than an OCD-specific characteristic[31,49]. Consistent with these reports, when comparing with healthy volunteers we found evidence for impaired acquisition of model-based control among patients with OCD, when learning directly and exclusively from experience. No difference was found in comparisons between healthy volunteers and individuals with other mood and anxiety disorders during uninstructed experience. Surprisingly, following debriefing we did not observe deficits in the ability of OCD participants to adopt a model-based strategy, demonstrating that, under some conditions, individuals with OCD recruit model-based control as readily as healthy volunteers, which is of particular interest given the established efficacy of cognitive-behavioural therapy (CBT) in the treatment of OCD[50]. We further observed a difference in the effect of debriefing between medicated and non-mediated OCD patients, with a reduction in the use of model-free control in medicated

but not non-medicated patients. Although it has been shown that, in OCD participants, CBT does not change use of model-based control in the original two-step task[51], our results suggest that pharmacological treatment may have an effect on the ability to suppress model-free control and modify behaviour once a correct model is acquired.

There are substantial differences between our paradigm and the original two-step task, that may explain the different pattern of deficits observed in individuals with OCD. Our task is structurally simpler due to having no choice at the second step, which will reduce working memory load and hence make model-based control easier. Indeed, while in the original two-step task working memory capacity is correlated with the use of a model-based strategy[22,23], we did not find any such correlations here, neither in healthy volunteers nor in clinical populations. The fact that participants in our task have extensive prior experience before being told its structure may also help them understand or remember this information when it is provided. We also found that participants who gave correct answers to pre-debriefing questionnaires showed a higher influence on model-based action values at session 4 (*See "Explicit reports about task structure are dissociated from uninstructed behaviour" in Supplementary Information*), providing further support to this idea. Furthermore, in our task, actions and states were differentiated by location rather than identity of visual stimuli, and these locations were fixed across trials rather than randomized as in the original two-step task. This allows model-free RL operating over spatial-motor representations, recently demonstrated in the original two-step task[52], to contribute more meaningfully to choice. Fixed spatial-motor contingencies also permit use of action-outcome, in addition to stimulus-outcome, mappings for model-based control, with the former thought to preferentially recruit the anterior cingulate cortex, rather than the orbitofrontal cortex[53,54]. An additional consequence of using fixed stimulus locations is that motor-systems can predict upcoming actions. Our observation of robust reaction-time differences following common vs rare transitions at session 1, when choices were model-free, suggests a dissociation between motor and cognitive systems in task structure learning, with motor systems learning to predict upcoming actions earlier and more readily than cognitive systems learn to use a model to guide choices, consistent with other recent reports[55]. Intriguingly, implicit and choice-based measures of task structure learning were correlated across participants, even at session 1, suggesting that this dissociation is only partial, with cognitive task models potentially informed by earlier motor-level learning. Finally, unlike the original two-step task, where model-based and model-free RL achieved similar reward rates[43,44], here use of model-based RL positively correlated with reward rate, generating a desirable trade-off between performance and cognitive effort that may influence arbitration between strategies[44].

In addition to increasing model-based control, debriefing had unexpected effects on model-free value updates, increasing the influence of second-step state values relative to trial outcomes on model-free first-step action values. This effect was robust and replicated in both clinical groups and healthy volunteers, including in the slow-paced task version. We hypothesized that this was mediated by debriefing modifying internal representations of the task state-space. Knowledge that the reward probability depended on the second-step state that was reached likely made internal representation of these states more salient and differentiated, and hence better able to accrue value, and thus driving model-free learning

at the first step. Consistent with this hypothesis, participants who became more strongly model-based following debriefing - indicating that they had acquired a correct model of the task, showed larger changes in model-free value updates. This result emphasizes that model-free RL operates over an internal representation of the states of the external world that must be learnt from sometimes ambiguous experience and is malleable in the face of new information[56]. A second unexpected effect of debriefing was an increased tendency to repeat choices, as indexed by the RL model's perseveration parameter. We hypothesized that this effect was mediated by explicit knowledge that the reward probabilities changed only occasionally. Consistent with this hypothesis, post-debriefing increases in perseveration correlated with decreases in perseveration over the course of each block. It thus seems that, after debriefing, participants inferred the occurrence of a reward probability reversal, and expected stability in the trials immediately following. These two unexpected findings show that the 'model' of a task that may be acquired through explicit information comprises not just the action-state contingencies that are required for model-based RL, but also beliefs about which distinct states of the environment are relevant for behaviour, and how the world may change over time, both of which can influence 'model-free' value learning.

It is also important to note that the most significant difference between clinical populations and healthy volunteers following debriefing was that, while the latter became more perseverative in their choices, this effect was smaller in OCD, not statistically significant in individuals with other mood and anxiety disorders, and we did not find significant correlations in either patient population with changes in perseveration from early to late in post-debriefing blocks. This evidence that inference based updating was impaired in OCD and other psychiatric diagnoses is particularly interesting given that the orbitofrontal cortex, which is consistently dysfunctional in OCD patients[57–59], is thought to build cognitive maps needed to infer task states that are not directly observable from sensory input[60].

We note limitations and directions for future studies. First, though analysing behaviour through the lens of model-based and model-free RL has yielded important insights, this dichotomy does not capture the full space of possible learning algorithms[12,33,61], and can obscure their dependence on common computational primitives such as a representation of the task state-space[62]. Although standard model-free and model-based algorithms provided a better fit to participants behaviour than other models tested, our exploration of possible models was necessarily not exhaustive, and we did not attempt to model learning the state-space itself, nor effects of instruction on this. Second, though we used several task variants, they were all adaptations of the original two-step task, and share with it both a comparatively small state space and probabilistic action-state transitions. It therefore remains an open question how broadly our findings generalise to other tasks. Model-based control may be more advantageous in larger state spaces, but model-learning and planning are correspondingly harder. Given our findings suggesting instruction shaped representation of the state-space, it would be interesting to explore instruction effects in tasks where there is ambiguity about the current state, or which state features are relevant for learning[63,64]. Another question is how information given to participants about their objectives shapes learning and use of task models. We told subject to 'gain as many rewards as possible' and it is possible that this focussed their attention on action-reward relationships to the detriment of action-state learning. This might explain why in an earlier study, participants were able

to successfully learn a task model during exposure to transition statistics in the absence of reward, then use it in a subsequent reward guided task[15]. Finally, it would be worth looking parametrically at the effect of instructions as a function of the amount of uninstructed task experience.

Regarding our findings in clinical populations, since prior work has identified that symptom dimensions can be a better predictor of behavioural phenotypes than clinical diagnoses, applying our task with dimensional methods in large online samples could provide further insight into clinical differences in learning and instruction effects[49]. Finally, although we know of no study showing a positive correlation between years of education and use of model-based or model-free control, the small but significant difference in terms of education between healthy volunteers and individuals with the mood and anxiety disorders might limit the comparisons between these samples. We also note that we observed substantial heterogeneity across participants in uninstructed behaviour, and it is likely that increased variability is an inherent feature of uninstructed tasks that may complicate assessing group differences.

In summary, we developed a sequential decision task which dissociates the effects of uninstructed experience and explicit information on RL strategy. We found that model-free RL dominates initial behaviour and maintains a strong influence throughout uninstructed learning, with model-based RL emerging only in a subset of individuals prior to receiving task structure information, and to a lesser extent in individuals with OCD. Receiving such information strongly increased model-based control, both in healthy individuals and those with OCD and other mood and anxiety diagnoses. Use of this task to dissociate effects of implicit and explicit information on RL strategy thus offers further insight into the content of learning and the imbalance between RL systems in neuropsychiatric disorders.

Methods

Participants and Testing Procedures

The research protocol was conducted in accordance with the declaration of Helsinki for human studies of the World Medical Association and approved by the Ethics Committees of the Champalimaud Centre for the Unknown, NOVA Medical School and Centro Hospitalar Psiquiátrico de Lisboa (CHPL), and the Institutional Review Board of the New York State Psychiatric Institute (NYSPI). Adult non-elderly participants (ages 18-65 years) were eligible and written informed consent was obtained from all prior to participation. Clinical samples were recruited at the Champalimaud Clinical Centre (CCC), CHPL and the NYSPI. In each of these centres, individuals with OCD were recruited from clinical or research databases. A mood and anxiety disorder control group was recruited randomly from patient lists (CCC and CHPL), or sequentially (NYSPI), among individuals with the following diagnoses: major depressive episode or disorder, dysthymia, bipolar disorder, generalized anxiety disorder, post-traumatic stress disorder, panic disorder or social anxiety disorder. Healthy controls were recruited sequentially as a convenience sample of community-dwelling participants and tested at the same locations. Participants were compensated for travel expenses plus a monetary bonus, ranging from 10€ to 25€ according

to performance in the task. These values were increased to 15$ and 35$ for the NY groups due to higher cost of living.

Following consent, each participant was screened for the presence of exclusion criteria using a clinical questionnaire assessing history of: acute medical illness; active neurological illness; clinically significant focal structural lesion of the central nervous system; history of chronic psychosis, dementia, developmental disorders with low intelligence quotient or any other form of cognitive impairment and illiteracy. Active psychiatric illness, including substance abuse or dependence, was also an exclusion criterion, with the exception of the diagnoses defining inclusion in the OCD and the mood and anxiety groups. In the absence of exclusion criteria, each participant then performed the simplified two-step task (see below).

Participants also performed a battery of structured interviews, scales and self-report inventories, including the MINI Neuropsychiatric Interview[65], the Structured Clinical Interview for the DSM-IV[66], the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS)[67,68] and the State-Trait Anxiety Inventory (STAI)[69]. As the groups recruited in Lisbon were assessed using the Y-BOCS-II while the groups recruited in New York were assessed using the original Y-BOCS, we converted the Y-BOCS-II score into original Y-BOCS score by transforming each item which was scored as 6 into a score of 5[68,70]. In the groups recruited in Lisbon, the Beck Depression Inventory-II (BDI-II)[71] was also applied to assess depressive symptoms, the Corsi block-tapping task to assess working memory[72] and a Go/No-Go task to assess impulsivity[73], while in New York, the Depression Anxiety Stress Scales (DASS)[74] were applied to assess symptoms of depression, anxiety and stress. Group differences in sociodemographic and psychometric measures were tested using one-way ANOVA for continuous variables (with Tukey's HSD for multiple comparisons) and Pearson's chi-squared for categorical variables. Correlations between neuropsychological test measures (Corsi; Go/No-GO) and the simplified two-step task measures were performed using Pearson's product moment correlation coefficient.

Simplified two-step task

The simplified two-step task was implemented in MATLAB R2014b using Psychtoolbox (Mathworks, Inc., Natick, Massachusetts, USA). The task consisted of a self-paced computer interface with 4 circles always visible on the screen: 2 central circles (upper and lower) flanked by two side circles (left and right) (Figure 1). Each circle was coloured yellow when available for selection, and black when unavailable, and could be selected by pressing the corresponding arrow key (up, down, left or right) on the computer keyboard. Each trial started with both of the central circles turning yellow, prompting a choice between the two (Figure 1a). This first step choice then activated one of the side circles in a probabilistic fashion, according to a structure of transition probabilities described below (Figure 1b). The active side circle could be selected with the corresponding arrow key, resulting either in reward (indicated by the circle changing to the image of a coin) or no reward (indicated by the circle changing to black). The reward probabilities on the right and left side changed in blocks that were either neutral (p=0.4 on each side) or non-neutral (p=0.8 on one side and p=0.2 on the other; Figure 1c). Changes from non-neutral blocks were triggered based on each subject's behaviour, occurring 20 trials after an exponential

moving average (tau = 8 trials) crossed a 75% correct threshold. In half of the cases this led to the other non-neutral block (reward probability reversals), and the other half to a neutral block. Changes from neutral blocks occurred with 10% probability on each trial after the 40th trial of that block, and always led to the non-neutral block that did not precede that neutral block. All participants performed 1200 trials on the same day, divided in 4 sessions of 300 trials each.

We ran two variants of the task which differed with respect to whether the transition probabilities linking the first-step actions to the second step states were fixed or underwent reversals. In both cases these probabilities were defined such that choosing one of the central circles (e.g. up) would cause one of the side circles (e.g. left) to turn yellow with high probability (p=0.8 – common transition), while causing the other side circle to turn yellow only in a minority of trials, i.e., with low probability (p=0.2 – rare transition). Choosing the other central circle would lead to common and rare transitions to the opposite sides. In the Fixed task, the transition probabilities were fixed for each individual throughout the entire task (e.g., common transitions for up-left and down-right, and rare transitions for up-right and down-left). In the Changing task, the transition probabilities underwent reversals on 50% of reward probability block changes after non-neutral blocks, such that the common transition became rare and vice versa (Figure 1b). In an initial group of healthy volunteers recruited in Lisbon, participants were randomized between the two versions of the task. In all clinical samples as well as healthy volunteers from New York, however, only the Fixed task was used.

Prior to starting the task, participants were given minimal information about task structure. They were only told that arrow keys could be used to interact with the screen, and that the image of a coin signalled accrual of a monetary reward. To test how providing explicit information about the task structure affected behaviour, debriefing was provided between the 3rd and the 4th sessions in some participants, with the 4th session of the task performed immediately after debriefing. Among healthy volunteers recruited in Lisbon and randomized between the two versions of the task, debriefing was performed in 17 of the 40 participants performing the Fixed version and in 16 of the 42 participants performing the Changing version of the task. In all other samples, debriefing was performed for everyone. Please see “Information provided to study participants” in *Supplementary information* for the specific information provided to participants prior to the task and during debriefing.

No statistical methods were used to pre-determine sample sizes but our sample sizes are [similar to/larger than] those reported in previous publications[11,17–25].

20 additional healthy volunteers performed a slower pace version of the task (with fixed transition probabilities). In this version, a 1-second delay was implemented after the first-step stimuli were shown, and before the participant could make the choice. The delay was signalled by having the upper and lower circle represented in pale yellow during the first second. After this time had elapsed, the circles turned bright yellow. A similar delay occurred at the second-step and a 1-second inter-trial interval was also implemented. In this version, participants performed three pre-debriefing sessions of 150 trials each and one post-debriefing session of 150 trials (total of 600 trials).

Europe PMC Funders Author Manuscripts

Data analysis

Data analysis was performed using Python version 3.7 (Python Software Foundation, http://python.org), SPSS (Version 21.0, SPSS Inc., Chicago, IL, USA), and R version 4.1.0 (R Core Team, https://www.R-project.org/). Except where noted otherwise, data are presented as mean (standard deviation).

Analysis of Stay Probability

The first analysis used to assess model-free vs. model-based behavioural strategies was an analysis of 'stay-probability'[11,17,26,27,18–25], defined as the probability of repeating the first-step choice on any given trial as a function of the outcome (rewarded or not) and transition (common or rare) on the previous trial. In addition to plotting raw stay probabilities, we quantified the effect of trial events on the subsequent choice using a logistic regression model, allowing other influences on choice such as participants biases and cross trial correlations (see below) to be taken into account. The *outcome*, *transition* and *transition-outcome interaction* predictors modelled the influence of the previous trial's outcome, transition and their interaction on the probability of repeating the previous first step choice. We additionally included a *bias* predictor capturing bias towards the upper or lower circle, and a *correct* predictor, which modelled the influence of whether the previous trials choice was correct (i.e. to the high reward probability option) on the probability of repeating that choice. The *correct* predictor prevents cross-trial correlations from generating spurious loading on the *transition-outcome interaction* predictor, which can occur in two-step tasks with high contrast between good and bad options, due to correlation between action values at the start of the trial and subsequent trial events[43].

Bootstrap tests were used to assess whether population mean predictor loadings in the logistic regression analysis were significantly different from zero. An ensemble of 5000 bootstrap resampled datasets were created by sampling participants from the original dataset with replacement. The logistic regression was run on each resampled dataset to estimate the sampling distribution of the population mean predictor loadings. The P value for predictor x was calculated based on this distribution as:

$$P = 2\min\left(\frac{\mathrm{M}}{N}, 1 - \frac{\mathrm{M}}{N}\right)$$

Where N is the total number of resampled datasets and M is the number of resampled datasets for which $x>0$.

When describing bootstrap test results we report the population mean predictor loading, the 95% confidence interval from the bootstrap resampled distribution, and the P value.

RL modelling

Additional analyses of behavioural strategy were obtained by fitting reinforcement learning models to observed behaviour. We first detail the model used for the main analyses then a set of alternative models that were rejected by model-comparison. The model followed those typically used in analysis of the original two-step task[11] in combining a model-based

and a model-free RL component, both with value estimates contributing to behaviour. The model-free component maintained estimates of the values $Q^{mf}(a)$ of the first-step actions (up or down), and $V(s)$ of the second step states (left and right). These values were updated as:

$$Q^{mf}_{t+1}(a) = (1-\alpha_Q)Q^{mf}_t(a) + \alpha_Q(\lambda r + (1-\lambda)V_t(s))$$

$$V_{t+1}(s) = (1-\alpha_Q)V_t(s) + \alpha_Q r$$

Where r is the reward obtained on trial t (1 or 0), α_Q is the value learning rate and λ is the eligibility trace parameter.

The model-based component maintained estimates of the transition probabilities linking the first step actions to the second step states ($P(s_2|a_1)$), updated as:

$$P_{t+1}(s|a) = (1-\alpha_T)P_t(s|a) + \alpha_T$$

$$P_{t+1}(s'|a) = (1-\alpha_T)P_t(s'|a)$$

where α_T is a learning rate for transition probabilities, s is the second step state reached and s' the second step state not reached on trial t.

At the start of each trial, model-based action values were calculated as:

$$Q^{mb}_t(a) = \sum_j P(s_j|a)Q_{mf}(s_j)$$

Model-free and model-based action values were combined with perseveration and bias to give net action values, calculated as:

$$Q^{net}_t(a_i) = G_{mf}Q^{mf}_t(a_i) + G_{mb}Q^{mb}_t(a) + bB_i + pP_i$$

Where G_{mf} and G_{mb} are parameters controlling, respectively, the strength of influence of model-free and model-based action values on choice. b is a parameter controlling the strength and direction of choice bias, b_i is a variable which takes a value of 1 for the up action and 0 for the down action. Positive values of b therefore generate a bias towards the up action and negative values towards the down action. p is a parameter controlling the strength and direction of choice perseveration, P_i is a variable which takes a value of 1 if action a_i was chosen on the previous trial and 0 if it was not. Positive values of p therefore promote repeating the previous choice while negative values promote switching.

The model's probability of choosing action a_i was given by $P(a_i) = \frac{{}_eQ^{net}(a_i)}{\sum_j {}_eQ^{net}(a_j)}$.

For model comparison, several reduced variants were considered. For the *Model-free only* variant the model-based component was removed such that the net action values were:

$$Q_t^{net}(a_i) = G_{mf}Q_t^{mf}(a_i) + bB_i + pP_i\,.$$

For the *Model-based only* variant the model-free component was removed such that the net action values were:

$$Q_t^{net}(a_i) = G_{mb}Q_t^{mb}(s_1, a_i) + bB_i + pP_i\,.$$

For the *No bias* variant the bias strength variable b was set to zero. For the *No perseveration* variant the perseveration strength variable p was set to zero.

We used separate weights (G_{mf}, G_{mb}) for the influence of the model-based and model-free systems[31], rather than tying them together as $G_{mf} = 1 - G_{mb}$ and using a separate softmax temperature parameter as in Daw et al. 2011[11].

As it has been proposed that apparently model-free behaviour in two-step tasks might in fact be generated by model-based strategies with incorrect beliefs[33], we additionally compared the goodness of fit for three such models.

The first was an 'unlucky symbol' model[33] which believed that one of the first step actions is unlucky and reduces the reward probability at trial outcome irrespective of which state is reached. To model this without making assumptions about which first step action was lucky and which was unlucky, we modified a standard model-based agent such that the first step action values for the *up* and *down* actions were given by:

$$Q_t^{mb}(up) = 2\mathrm{L}\sum_j P(s_j|up)Q_{mf}(s_j)$$

$$Q_t^{mb}(down) = 2(1-\mathrm{L})\sum_j P(s_j|down)Q_{mf}(s_j)$$

Where L is a parameter which determines how lucky one first step action is considered relative to the other, constrained to lie on the range 0-1. When $L = 0.5$ the model is identical to a standard model-based agent, as L approaches 0 or 1 the values for the two actions are scaled relative to each other.

The second incorrect model considered was the 'transition dependent learning rate' model of Silva and Hare[33], which has a different learning rate for the value of the second-step action depending on whether the preceding transition was common or rare. We modelled this by

adapting a standard model-based agent to use separate parameters for the value learning rate at the second step following transitions estimated to be common or rare based on the agents current beliefs about the transition probabilities.

The third incorrect model was motivated by a reviewer's suggestion that participants apparently model-free behaviour might reflect a belief that state transitions were deterministic but highly volatile, such that only the most recently observed transition for a given first-step action is informative about its transition probabilities. We modelled this using a version of the standard model-based agent in which the learning rate for transitions α_T was fixed at 1.

Hierarchical modelling:

Fits of both the logistic regression model and reinforcement learning models to populations of participants used a Bayesian hierarchical modelling framework [75], in which parameter vectors h_i for individual sessions were assumed to be drawn from Gaussian distributions at the population level with means and variance $\theta = \{\mu, \Sigma\}$. The population level prior distributions were fit to their maximum likelihood estimate:

$$\theta^{ML} = argmax_\theta \left\{ p(D|\theta) = argmax_\theta \left\{ \prod_i^N \int dh_i \, p(D_i|h_i) p(h_i|\theta) \right\} \right.$$

Optimization was performed using the Expectation-Maximization algorithm with a Laplace approximation for the E-step at the k-th iteration given by:

$$p\left(h_i^k \middle| D_i\right) = N\left(m_i^k, V_i^k\right)$$

$$m_i^k = argmax_h \left\{ p(D_i|h) p\left(h \middle| \theta^{k-1}\right) \right\}$$

Where $N\left(m_i^k, V_i^k\right)$ is a normal distribution with mean m_i^k given by the maximum a posteriori value of the session parameter vector h_i given the population level means and variance θ^{k-1}, and the covariance $\mathbf{V}_i^k$ given by the inverse Hessian of the likelihood around m_i^k. For simplicity we assumed that the population level covariance Σ had zero off-diagonal terms. For the k-th M-step of the EM algorithm the population level prior distribution parameters $\theta = \{\mu, \Sigma\}$ are updated as:

$$\mu^k = \frac{1}{N} \sum_{i=1}^{N} m_i^k$$

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} \left[\left(m_i^k\right)^2 + V_i^k \right] - \left(\mu^k\right)^2$$

Parameters were transformed before inference to enforce constraints:

$$0 < \{ G_{mf}, G_{mb} \}$$

$$0 < \{ \alpha_Q, \alpha_T, \lambda \} < 1$$

95% confidence intervals on population means μ were calculated as $c_i = \pm 1.96\sqrt{-1/H_i}$ where c_i is the confidence interval for parameter i and H_i is the i-th diagonal element of the Hessian at θ^{ML} with respect to μ.

Parameter recovery test

To test the accuracy with which model parameters could be recovered from simulated data (Supplementary Figure 7), we fit the hierarchical RL model to a given behavioral dataset to obtain the means and variances (μ, Σ) of the population level distributions for each parameter. Then for for each parameter P, we generated a set of simulated datasets where all parameters except P were drawn randomly for each subject from the fitted population level distributions, and P was systematically varied across the range of parameters values, e.g. for unit range parameters P was varied between 0.1 and 0.9 in steps of 0.1. For each value of P, 10 simulated datasets were generated, Supplementary Figure 7 shows the mean and standard deviation across these repeats.

Model comparison

To compare the goodness of fit for hierarchical models with different numbers of parameters we used the integrated Bayes Information Criterion (iBIC) score. The iBIC score is related to the model log likelihood $p(D|M)$ as:

$$\log p(D|M) = \int d\theta \, p(D|\theta) p(\theta|M)$$

$$\approx -\frac{1}{2} iBIC = \log p\left(D \middle| \theta^{ML}\right) - \frac{1}{2}|M|\log|D|$$

Where |M| is the number of fitted parameters of the prior, |D| is the number of data points (total choices made by all participants) and iBIC is the integrated BIC score. The log data likelihood given maximum likelihood parameters for the prior $\log p\left(D \middle| \theta^{ML}\right)$ is calculated by integrating out the individual session parameters:

$$\log p\left(D \middle| \theta^{ML}\right) = \sum_i^N log \int dh \, p(D_i|h) p(h \middle| \theta^{ML})$$

$$\approx \sum_{i}^{N} log \frac{1}{K} \sum_{j=1}^{K} p\left(D_i \middle| h^j\right)$$

Where the integral is approximated as the average over K samples drawn from the prior $p(h|\theta^{ML})$.

Permutation tests

Permutation testing was used to assess the statistical significance of learning and instruction effects on RL and logistic regression model fits, and the reversal analysis. To assess the effects of experience in the task we compared behaviour between sessions 1 and 3, while to assess the effects of explicit knowledge we compared behaviour between sessions 3 and 4 in the groups that did and did not receive instruction between these sessions.

To test for a significant difference in behavioural parameter x between conditions (e.g. session 1 vs 3), we evaluated the population mean value of the parameter for each conditions and calculated the difference Δx_{true} between them. We then constructed an ensemble of 5000 permuted datasets in which the assigments of sessions to the two conditions was randomised. Randomisation was performed within subject, such that the number of sessions from each subject in each condition was preserved. For each permuted dataset we re-ran the analysis and evaluated the difference in parameter x between the two conditions, to give a distribution of Δx_{perm}, which in the limit of many permutations is the distribution of Δ_x under the null hypothesis that there is no difference between the conditions. The two tailed P value for the observed difference is given by:

$$P = 2 \min\left(\frac{M}{N}, 1 - \frac{M}{N}\right)$$

Where N is the number of permutations and M is the number of permutations for which $\Delta x_{perm} > \Delta x_{true}$.

To assess significant differences in learning or debriefing effects between clinical groups and healthy controls, and for differences in the healthy controls between groups who did and did not receive debriefing, we tested for a significant interaction between session number and group. The significance of the interaction was assessed using a permutation test in which we evaluated the difference $\Delta g_{true} = \Delta x^A_{i,j} - \Delta x^B_{i,j}$ where $\Delta x^A_{i,j}$ is the difference in behavioural parameter x between sessions i and j in group A, and $\Delta x^B_{i,j}$ is the difference in behavioural parameter x between sessions i and j in group B. We then constructed an ensemble of 5000 permuted datasets by randomly permuting participants between groups while preserving the total number of participants in each group. We assessed $\Delta g_{perm} = \Delta x^A_{i,j} - \Delta x^B_{i,j}$ for each permuted dataset and calculated P values for the interaction as above.

Europe PMC Funders Author Manuscripts

When reporting permutation test results we indicate the observed difference in the data, the 95% confidence interval for this difference under the null hypothesis (null 95% CI) and the P value.

Other statistical analyses

To explore further differences between groups, we used one-way ANOVA's with Tukey's HSD; Pearson's Chi-squared; Sign test; Paired t-test; repeated measures ANOVA and mixed ANOVA. The decision on whether to use a parametric or a non-parametric test was based on visual inspection of the distribution and on the central limit theorem. All tests were two-tailed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

PCR was supported by a doctoral fellowship (reference SFRH/SINTD/94350/2013) from Fundação para a Ciência e Tecnologia and by a Fulbright Research Grant from Bureau of Educational and Cultural Affairs of the US Department of State. T.A. was funded by was funded by Wellcome Trust grants WT096193AIA, 202831/Z/16/Z and 214314/Z/18/Z. AM is supported by a doctoral fellowship (reference SFRH/BD/144508/2019) from Fundação para a Ciência e Tecnologia. JBBC is supported by grant PTDC/MEC-PSQ/30302/2017-IC&DT-LISBOA-01-0145-FEDER, funded by national funds from FCT/MCTES and co-funded by FEDER, under the Partnership Agreement Lisboa 2020 - Programa Operacional Regional de Lisboa. PD is supported by the Max-Planck-Gesellschaft (Max Planck Society) and the Alexander von Humboldt-Stiftung (Alexander von Humboldt Foundation). AJOM is supported by grant PTDC/MEC-PSQ/30302/2017-IC&DT-LISBOA-01-0145-FEDER, funded by national funds from FCT/MCTES and co-funded by FEDER, under the Partnership Agreement Lisboa 2020 - Programa Operacional Regional de Lisboa, by grants PTDC/MED-NEU/31331/2017 and PTDC/MEC-PSQ/30302/2017 from Fundação para a Ciência e Tecnologia, and by a Starting Grant from the European Research Council (ERC) under the European

Union's Horizon 2020 research and innovation programme (grant agreement No 950357). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Data availability

The data used in the study is available from:

https://github.com/ThomasAkam/Two-step_explicit_knowledge

Code availability

Two-step task analysis code is available from:

https://github.com/ThomasAkam/Two-step_explicit_knowledge

References

1. Dickinson A. Actions and Habits: The Development of Behavioural Autonomy. Philos Trans R Soc B Biol Sci. 1985; 308: 67–78.
2. Sloman SA. The empirical case for two systems of reasoning. Psychol Bull. 1996; 119: 3–22.
3. Kahneman D. A perspective on judgment and choice: Mapping bounded rationality. Behav Sci. 2003; 58: 697–720.

Europe PMC Funders Author Manuscripts

4. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8: 1704–11. [PubMed: 16286932]
5. Dolan RJ, Dayan P. Goals and habits in the brain. Neuron. 2013; 80: 312–325. [PubMed: 24139036]
6. Robbins TW, Costa RM. Habits. Curr Biol. 2017; 27: R1200–R1206. [PubMed: 29161553]
7. Adams CD, Dickinson A. Instrumental responding following reinforcer devaluation. Q J Exp Psychol Sect B Comp Physiol Psychol. 1981; 33: 109–121.
8. Adams CD. Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q J Exp Psychol Sect B. 1982; 34: 77–98.
9. Colwill RM, Rescorla RA. Postconditioning Devaluation of a Reinforcer Affects Instrumental Responding. J Exp Psychol Anim Behav Process. 1985; 11: 120–132.
10. Sutton, RS, Barto, AG. Introduction to Reinforcement Learning. Vol. 4. The MIT press; 1998.
11. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69: 1204–1215. [PubMed: 21435563]
12. Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol. 2017; doi: 10.1371/journal.pcbi.1005768
13. Wan Lee S, Shimojo S, O'Doherty JP. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. Neuron. 2014; 81: 687–699. [PubMed: 24507199]
14. Gershman SJ, Horvitz EJ, Tenenbaum JB. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science. 2015; 349: 273–278. [PubMed: 26185246]
15. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010; 66: 585–595. [PubMed: 20510862]
16. Wunderlich K, Dayan P, Dolan RJ. Mapping value based planning and extensively trained choice in the human brain. Nat Neurosci. 2012; 15: 786–791. [PubMed: 22406551]
17. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci. 2013; 110: 20941–20946. [PubMed: 24324166]
18. Worbe Y, et al. Valence-dependent influence of serotonin depletion on model-based choice strategy. Mol Psychiatry. 2016; 21: 624–629. [PubMed: 25869808]
19. Friedel E, et al. Devaluation and sequential decisions: linking goal-directed and model-based behavior. Front Hum Neurosci. 2014; 8
20. Otto AR, Gershman SJ, Markman AB, Daw ND. The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. Psychol Sci. 2013; 24: 751–761. [PubMed: 23558545]
21. Skatova A, Chan PA, Daw ND. Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. Front Hum Neurosci. 2013; 7
22. Eppinger B, Walter M, Heekeren HR, Li SC. Of goals and habits: Age-related and individual differences in goal-directed decision-making. Front Neurosci. 2013; doi: 10.3389/fnins.2013.00253
23. Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. Neuron. 2013; 80: 914–919. [PubMed: 24206669]
24. Schad DJ, et al. Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. Front Psychol. 2014; 5
25. Radenbach C, et al. The interaction of acute and chronic stress impairs model-based behavioral control. Psychoneuroendocrinology. 2015; 53: 268–280. [PubMed: 25662093]
26. Deserno L, et al. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc Natl Acad Sci. 2015; 112: 1595–1600. [PubMed: 25605941]
27. Economides M, Kurth-Nelson Z, Lübbert A, Guitart-Masip M, Dolan RJ. Model-Based Reasoning in Humans Becomes Automatic with Training. PLoS Comput Biol. 2015; 11
28. Sebold M, et al. Model-based and model-free decisions in alcohol dependence. Neuropsychobiology. 2014; 70: 122–131. [PubMed: 25359492]

29. Voon V, et al. Disorders of compulsivity: a common bias towards learning habits. Mol Psychiatry. 2015; 20: 345–352. [PubMed: 24840709]
30. Voon V, et al. Motivation and value influences in the relative balance of goal-directed and habitual behaviours in obsessive-compulsive disorder. Transl Psychiatry. 2015; 5 e670 [PubMed: 26529423]
31. Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. Characterizing a psychiatric symptom dimension related to deficits in goaldirected control. Elife. 2016; 5
32. Culbreth AJ, Westbrook A, Daw ND, Botvinick M, Barch DM. Reduced model-based decision-making in schizophrenia. J Abnorm Psychol. 2016; 125: 777–787. [PubMed: 27175984]
33. da Silva CF, Hare T. Humans primarily use model-based inference in the two-stage task. Nat Hum Behav. 2020; 4 (10) 1053–1066. [PubMed: 32632333]
34. Kaufman, Arnold; Baron, Alan; Kopp, RE. Some Effects of Instructions on Human Operant Behavior. Psychon Monogr Suppl. 1966; 1: 243–50.
35. Baron A, Kaufman A, Stauber KA. Effects of instructions and reinforcement-feedback on human operant behavior maintained by fixed-interval reinforcement1. J Exp Anal Behav. 1969; doi: 10.1901/jeab.1969.12-701
36. Baron A, Galizio M. Instructional control of human operant behavior. Psychol Rec. 1983; 33 (4) 495.
37. Wilson GD. Reversal of Differential GSR Conditioning by Instructions. J Exp Psychol. 1968; 76: 491–93. [PubMed: 5642166]
38. Atlas LY, Doll BB, Li J, Daw ND, Phelps EA. Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. Elife. 2016; doi: 10.7554/elife.15192
39. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ. Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. Brain Res. 2009; 1299: 74–94. [PubMed: 19595993]
40. Biele G, Rieskamp J, Gonzalez R. Computational models for the combination of advice and individual learning. Cogn Sci. 2009; doi: 10.1111/j.1551-6709.2009.01010.x
41. Li J, Delgado MR, Phelps EA. How instructed knowledge modulates the neural systems of reward learning. Proc Natl Acad Sci. 2011; doi: 10.1073/pnas.1014938108
42. Hertwig R, Erev I. The description-experience gap in risky choice. Trends in Cognitive Sciences. 2009; doi: 10.1016/j.tics.2009.09.004
43. Akam T, Costa R, Dayan P. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. PLoS Comput Biol. 2015; 11
44. Kool W, Cushman FA, Gershman SJ. When Does Model-Based Control Pay Off? PLoS Comput Biol. 2016; 12
45. Balleine BW, Dickinson A. Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998; 37: 407–419. [PubMed: 9704982]
46. Bostan AC, Strick PL. The basal ganglia and the cerebellum: nodes in an integrated network. Nature Reviews Neuroscience. 2018; 1–13. DOI: 10.1038/s41583-018-0002-7
47. Thorndike EL. Animal intelligence: An experimental study of the associative processes in animals. Psychol Rev. 1898; 2: 1–107.
48. Biele G, Rieskamp J, Krugel LK, Heekeren HR. The Neural basis of following advice. PLoS Biol. 2011; doi: 10.1371/journal.pbio.1001089
49. Gillan CM, et al. Comparison of the Association between Goal-Directed Planning and Self-reported Compulsivity vs Obsessive-Compulsive Disorder Diagnosis. JAMA Psychiatry. 2020; doi: 10.1001/jamapsychiatry.2019.2998
50. Hirschtritt ME, Bloch MH, Mathews CA. Obsessive-compulsive disorder advances in diagnosis and treatment. JAMA - Journal of the American Medical Association. 2017; doi: 10.1001/jama.2017.2200
51. Wheaton MG, Gillan CM, Simpson HB. Does cognitive-behavioral therapy affect goal-directed planning in obsessive-compulsive disorder? Psychiatry Res. 2019; doi: 10.1016/j.psychres.2018.12.079

52. Shahar N, et al. Credit assignment to state-independent task representations and its relationship with model-based decision making. Proc Natl Acad Sci. 2019; doi: 10.1073/pnas.1821647116
53. Rushworth MFS, Behrens TEJ, Rudebeck PH, Walton ME. Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. Trends in Cognitive Sciences. 2007; doi: 10.1016/j.tics.2007.01.004
54. Akam T, Rodrigues-Vaz I, Marcelo I, Zhang X, Pereira Michael, Oliveira RF, Dayan P, Costa RM, Akam T. The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection. Neuron. 2021; 109: 149–163. [PubMed: 33152266]
55. Konovalov, Arkady; Krajbich, I. Mouse tracking reveals structure knowledge in the absence of model-based choice. Nat Commun. 2020; 11
56. Gershman SJ, Uchida N. Believing in dopamine. Nat Rev Neurosci. 2019; doi: 10.1038/s41583-019-0220-7
57. Baxter LR Jr, et al. Local cerebral glucose metabolic rates in obsessive-compulsive disorder. A comparison with rates in unipolar depression and in normal controls. Arch Gen Psychiatry. 1987; 44: 211–218. [PubMed: 3493749]
58. Menzies L, et al. Integrating evidence from neuroimaging and neuropsychological studies of obsessive-compulsive disorder: The orbitofronto-striatal model revisited. Neuroscience and Biobehavioral Reviews. 2008; 32: 525–549. [PubMed: 18061263]
59. Chamberlain SR, et al. Orbitofrontal dysfunction in patients with obsessive-compulsive disorder and their unaffected relatives. Science. 2008; 80 doi: 10.1126/science.1154433
60. Schuck NW, Cai MB, Wilson RC, Niv Y. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron. 2016; doi: 10.1016/j.neuron.2016.08.019
61. Piray, Payam; Daw, N. Linear reinforcement learning: Flexible reuse of computation in planning, grid fields, and cognitive control. Nature Communications. 2021; 12 (1) 1–20.
62. Collins AGE, Cockburn J. Beyond dichotomies in reinforcement learning. Nat Rev Neurosci. 2020; doi: 10.1038/s41583-020-0355-6
63. Farashahi S, Rowe K, Aslami Z, Lee D, Soltani A. Feature-based learning improves adaptability without compromising precision. Nat Commun. 2017; doi: 10.1038/s41467-017-01874-w
64. Farashahi S, Xu J, Wu SW, Soltani A. Learning arbitrary stimulus-reward associations for naturalistic stimuli involves transition from learning about features to learning about objects. Cognition. 2020; doi: 10.1016/j.cognition.2020.104425
65. Sheehan DV, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. Eur Psychiatry. 1997; 12: 232–241.
66. First MB, Spitzer RL, Gibbon M, Williams JBW. Structured Clinical Interview for DSM-IV Axis I Disorders. New York State Psychiatric Institute. 2002.
67. Goodman WK, et al. The Yale-Brown Obsessive Compulsive Scale: I. Development, Use, and Reliability. Arch Gen Psychiatry. 1989; 46: 1006–1011. [PubMed: 2684084]
68. Storch EA, et al. Development and psychometric evaluation of the yale-brown obsessive-compulsive scale-second edition. Psychol Assess. 2010; 22: 223–232. [PubMed: 20528050]
69. Spielberger C. Manual for the State-Trait Anxiety Inventory (STAI). Consult Psychol Press. 1983.
70. Castro-Rodrigues P, et al. Criterion validity of the Yale-Brown Obsessive-Compulsive Scale Second Edition for diagnosis of obsessive-compulsive disorder in adults. Front Psychiatry. 2018; doi: 10.3389/fpsyt.2018.00397
71. Beck AT, Steer RA, Brown GK. Manual for the Beck depression inventory-II. San Antonio, TX Psychol Corp. 1996. 1–82.
72. Berch DB, Krikorian R, Huha EM. The Corsi block-tapping task: methodological and theoretical considerations. Brain Cogn. 1998; 38: 317–38. [PubMed: 9841789]
73. Mueller ST, Piper BJ. The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. J Neurosci Methods. 2014; 222: 250–259. [PubMed: 24269254]
74. Lovibond SH, Lovibond PF. Manual for the Depression Anxiety Stress Scales. Psychology Foundation of Australia. 1995; doi: 10.1016/0005-7967(94)00075-U
75. Huys QJM, et al. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput Biol. 2011; 7

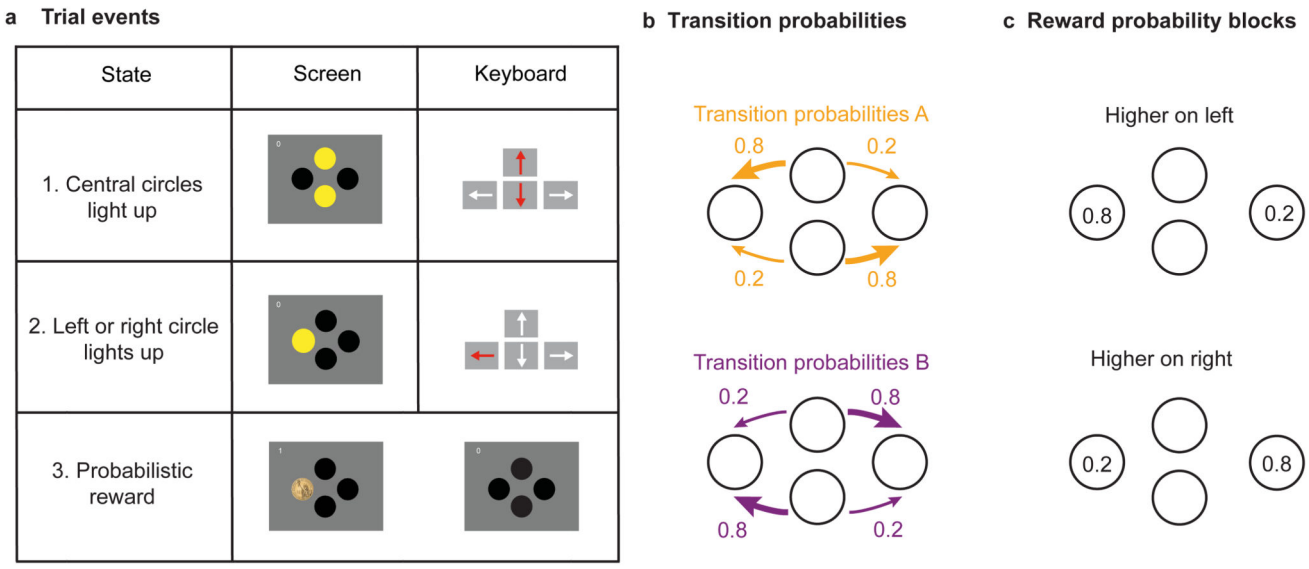

Figure 1. Behavioural Task.

The simplified two-step task was presented on a computer screen with 4 circles visible on a grey background: 2 central circles (upper and lower) and two side circles (right and left). Each circle was coloured yellow when available for selection, and black when unavailable. Circles could be selected by pressing the corresponding arrow key on the computer keyboard. **a)** Trial events. Each trial started with the central circles turning yellow, prompting the first-step choice between either upper or lower circle (a1). Following this, one of the side circles (left or right) would turn yellow (a2), with differing probabilities (see b). The subject then selects the yellow side circle resulting in a probabilistic monetary reward, indicated by the circle changing to the image of a coin (a3 left). No reward was indicated by the circle changing back to black (a3 right). **b)** Transition probabilities linking first step choice (up or down) to second-step state (left or right). Each first step option commonly (80% of trials) led to one second-step state and rarely (20% of trials) to the other. In the Fixed version of the task transition probabilities were counterbalanced across participants, with half experiencing the type A probabilities (top) and half the type B (bottom). In the Changing version of the task, the transition probabilities alternated between type A and B in blocks. **c)** Reward probability blocks. The reward probabilities for the side circles changed in blocks that were either higher on one or other sides (p=0.8 vs p=0.2, non-neutral blocks) or neutral (p=0.4 for both sides). Non-neutral blocks ended when participants consistently chose the first-step option that most frequently led to the high reward probability side. Neutral blocks ended probabilistically, independent of participants' behaviour (see methods). To maximize reward rate, participants must choose the first step action which commonly leads to the second-step state with higher reward probability, tracking the best option across reward-probability reversals.

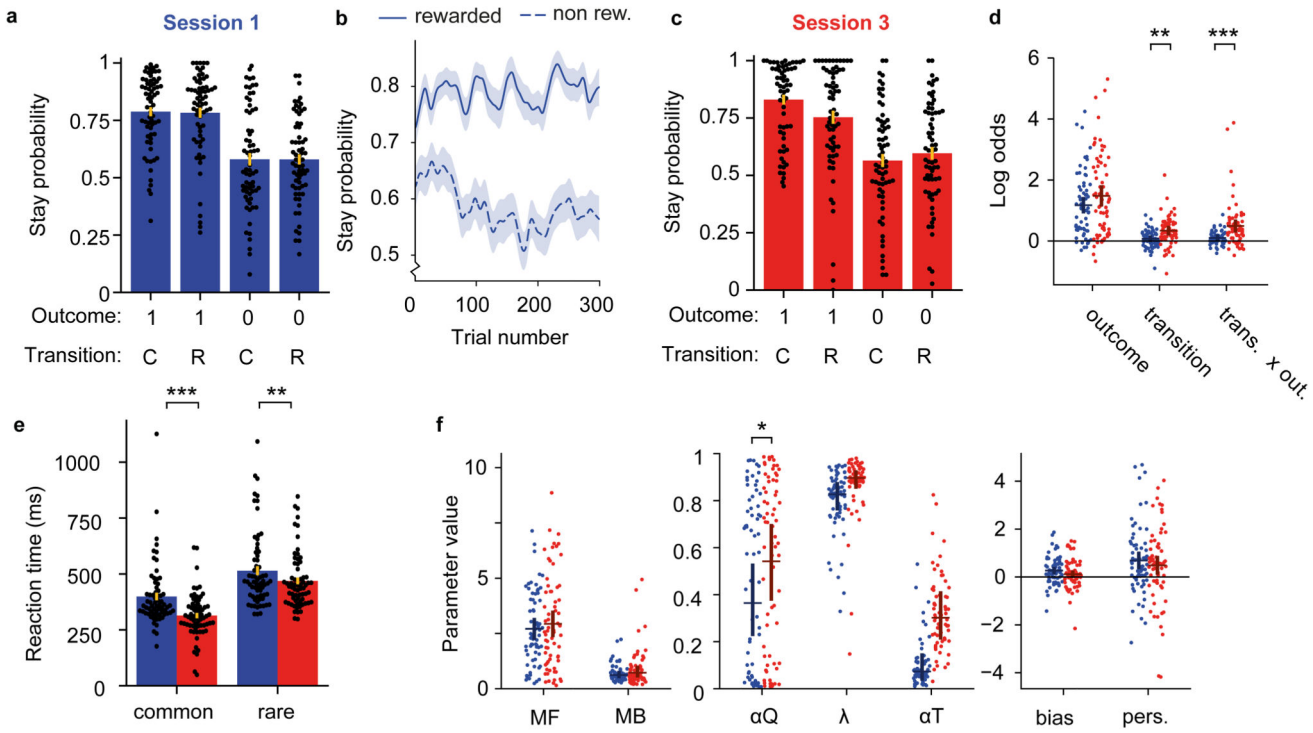

Figure 2. Uninstructed behaviour is predominantly model-free.

Analysis of uninstructed behaviour in 67 healthy volunteers. **a)** Session 1 stay probability analysis showing the probability of repeating the first step choice on the next trial as a function of trial outcome (rewarded or not rewarded) and state transition (common or rare). Error bars indicate cross subject standard error (SEM). **b)** Stay probability for rewarded and non-rewarded trials as a function of trial number in session 1. Shaded area shows across subject standard error. **c)** Stay probabilities for session 3. **d)** Logistic regression analysis of how the outcome (rewarded or not), transition (common or rare) and their interaction, predict the probability of repeating the same choice on the subsequent trial. Dots indicate maximum a posteriori parameter values for individual participants, bars indicate the population mean and 95% confidence interval of the mean. In this and other panels, blue indicates session 1 while red indicates session 3. The influence of both state transition (null 95% CI [-0.18,0.18], coefficient change=0.27, P=0.003, permutation test), and transition-outcome interaction (null 95% CI [-0.25,0.24], coefficient change=0.39, P<0.001) increased between session 1 and 3. **e)** Reaction times after common and rare transitions in session 1 and 3. Key-press reaction times at the second-step became faster overall between session 1 and 3 (main effect of session $F_{1,66}$=21.1, P<0.0001, η_p^2=0.24), but this was more pronounced following common than rare transitions (session-transition interaction $F_{1,66}$=21.1, P=0.008, η_p^2=0.1, repeated measures ANOVA). **f)** Comparison of mixture model fits between session 1 and session 3. Dots and bars are represented as in panel C. The value learning rate increased significantly between session 1 and 3 (null 95% CI [-0.17,0.17], parameter change=0.18, P=0.03). RL model parameters: MF: Model-free strength, MB: Model-based strength, αQ: Value learning rate, λ: Eligibility trace, αT:

Europe PMC Funders Author Manuscripts

Transition prob. learning rate, bias: Choice bias, pers.: Choice perseveration. In all figures significant differences are indicated as: * $P<0.05$, ** $P<0.01$, *** $P<0.001$.

Europe PMC Funders Author Manuscripts

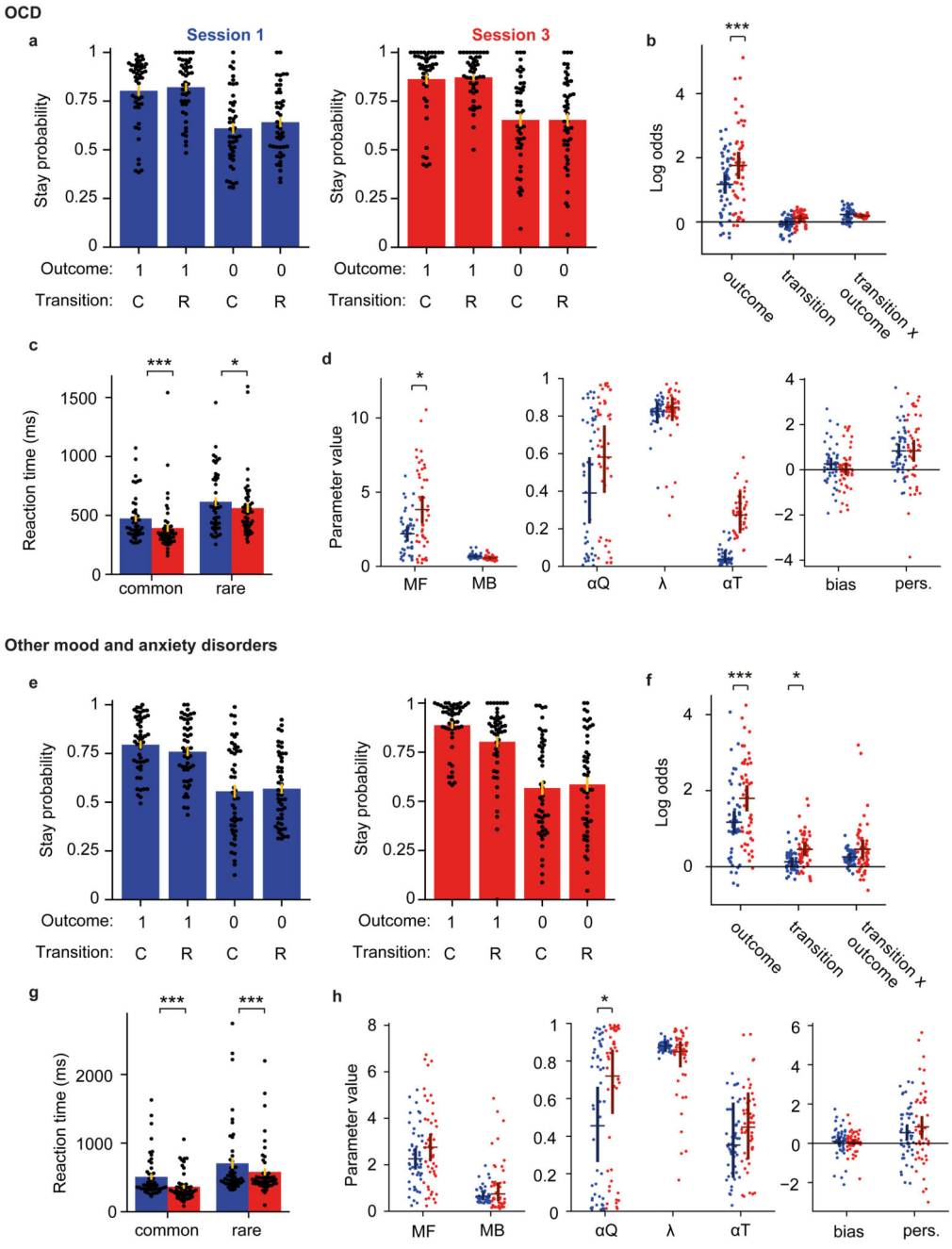

Figure 3. Impaired learning of model-based control from experience in OCD

Participants with OCD (n=46) are represented in panels **a-c**, those with mood and anxiety disorders (n=49) in panels **d-f**. **(a, e)** Stay probability analysis for session 1 (left, blue) and session 3 (right, red), as figure 2a. **(b, f)** Logistic regression analysis of stay probabilities, as figure 2d. In the OCD group the influence of trial outcome on stay probability increased between session 1 and 3 (null 95% CI [-0.35,0.36], coefficient change=0.58, P<0.001, permutation test). In the group with mood and anxiety disorders, the influence of outcome (null 95% CI [-0.33,0.35], coefficient change=0.63, P<0.001) and transition (null 95% CI

Europe PMC Funders Author Manuscripts

[-0.25,0.25], coefficient change=0.33, P=0.011) increased. **(c, g)** Second-step reaction times after common and rare transitions in session 1 and 3. In the OCD group, reaction times were faster following common than rare transitions (main effect of transition, F1,45=51.3, P<0.0001, ηp2=0.53, repeated measures ANOVA), and also in session 3 than session 1 (main effect of session, $F_{1,45}$=10, P=0.003, η_p^2=0.18). In the group with mood and anxiety disorders, second-step reaction times were faster following common than rare transitions (main effect of transition, $F_{1,48}$=34.2, P<0.0001, η_p^2=0.42, repeated measures ANOVA) and faster in session 3 than session 1 (main effect of session, $F_{1,48}$=30.5, P<0.0001, η_p^2=0.39). **(d, h)** Comparison of RL mixture model fits, as figure 2f. In the OCD group, the influence of model-free action values on choice increased between session 1 and 3 (null 95% CI [-1.36,1.32], parameter change=1.71, P=0.012, permutation test). In the mood and anxiety disorders group, the value learning rate increased between session 1 and 3 (null 95% CI [-0.21,0.21], coefficient change=0.26, P=0.011).

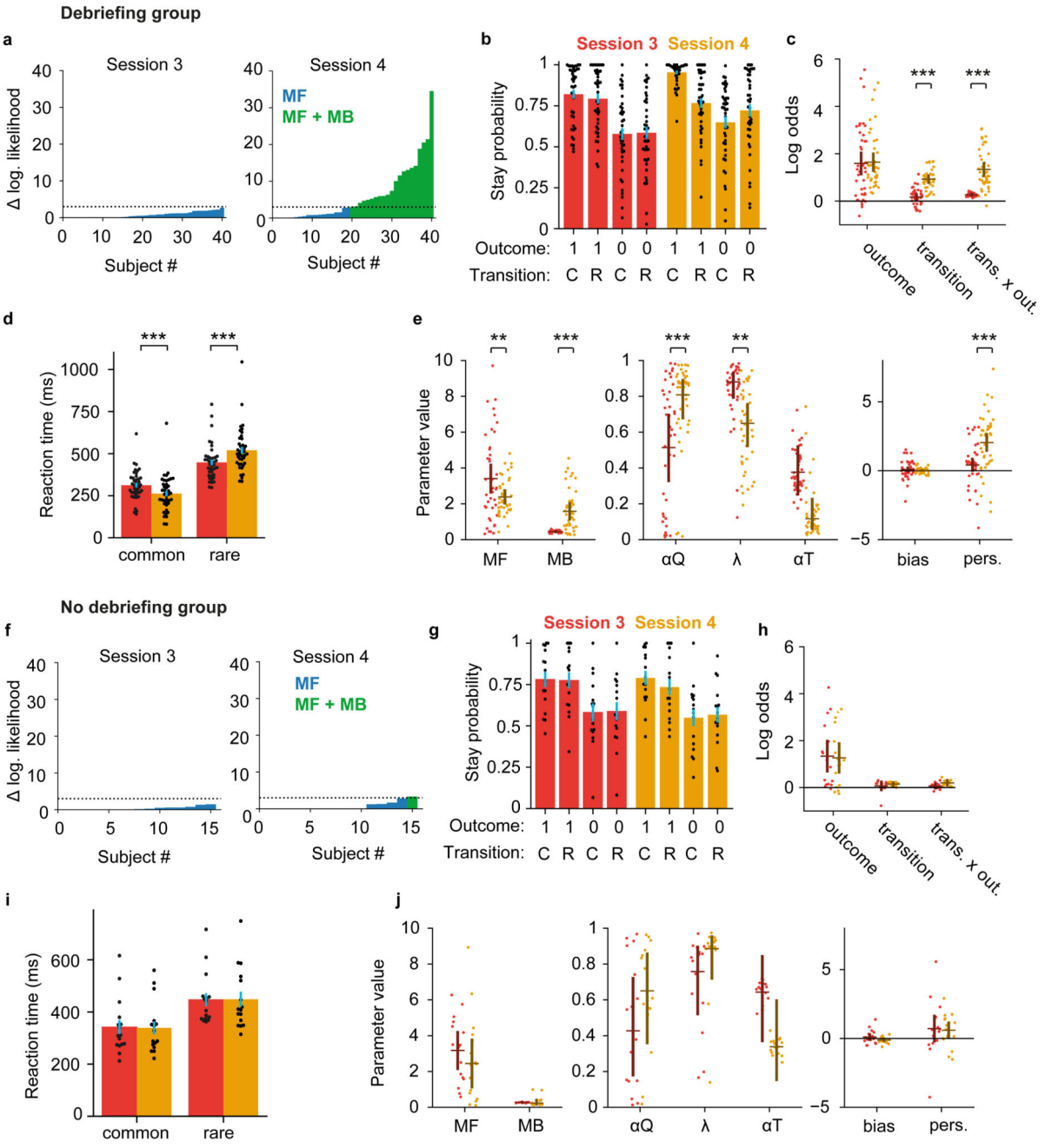

Figure 4. Explicit knowledge increases model-based control.

To avoid ceiling effects, analysis of debriefing effects was assess in healthy volunteers who were model-free at session 3, as assessed by likelihood ratio test (debriefing group n=41, **a-e**, no-debriefing group n=16, **f-j**). **(a, f)** Per-subject likelihood ratio test for use of model-based strategy at session 3 (left panel) and session 4 (right panel). Colour indicates whether each participant's data was better explained by a mixture of model-free and model-based RL (green) or model-free RL only (blue), using a $p<0.05$ threshold for rejecting the simpler model. Y-axis shows difference in log likelihood between the models. **(b,**

Europe PMC Funders Author Manuscripts

g) Stay probability analysis showing the probability of repeating the first-step choice on the next trial as a function of trial outcome and state transition. In these and remaining panels, red indicates session 3 while yellow indicates session 4. Error bars show cross subject standard error of the mean (SEM). **(c, h)** Logistic regression analysis of how the outcome, transition and their interaction, predict the probability of repeating the same choice on the subsequent trial. Dots indicate maximum a posteriori values for individual participants, while bars indicate the population mean and 95% confidence interval on the mean. Following debriefing the influence of state transition (null 95% CI [-0.42,0.42], coefficient change=0.75, P<0.001; permutation test) and transition-outcome interaction (95% CI [-0.51,0.50], coefficient change=1.07, P<0.001) increased. **(d, i)** Second-step reaction times following common and rare transitions. Following debriefing the influence of transition on reaction time increased (session-transition interaction F1,40=59.6, P<0.0001, η_p^2=0.59, repeated measures ANOVA). **(e, j)** Comparison of mixture model fits. Dots and bars are as in panel c. Following debriefing, the influence of model-based action values on choice increased (null 95% CI [-0.70,0.70], parameter change=1.17, P<0.001), the influence of model-free action values on choice decreased (null 95% CI [-0.79,0.79], parameter change=-1.04, P=0.006), value learning rate increased (null 95% CI [-0.18,0.18], parameter change=0.29, P<0.001), the eligibility trace parameter decreased (null 95% CI [-0.16,0.17], parameter change=-0.23, *P*=0.006) and the perseveration parameter increased (null 95% CI [-0.75,0.76], parameter change=1.63, *P*<0.001). RL model parameters as Fig. 2f.

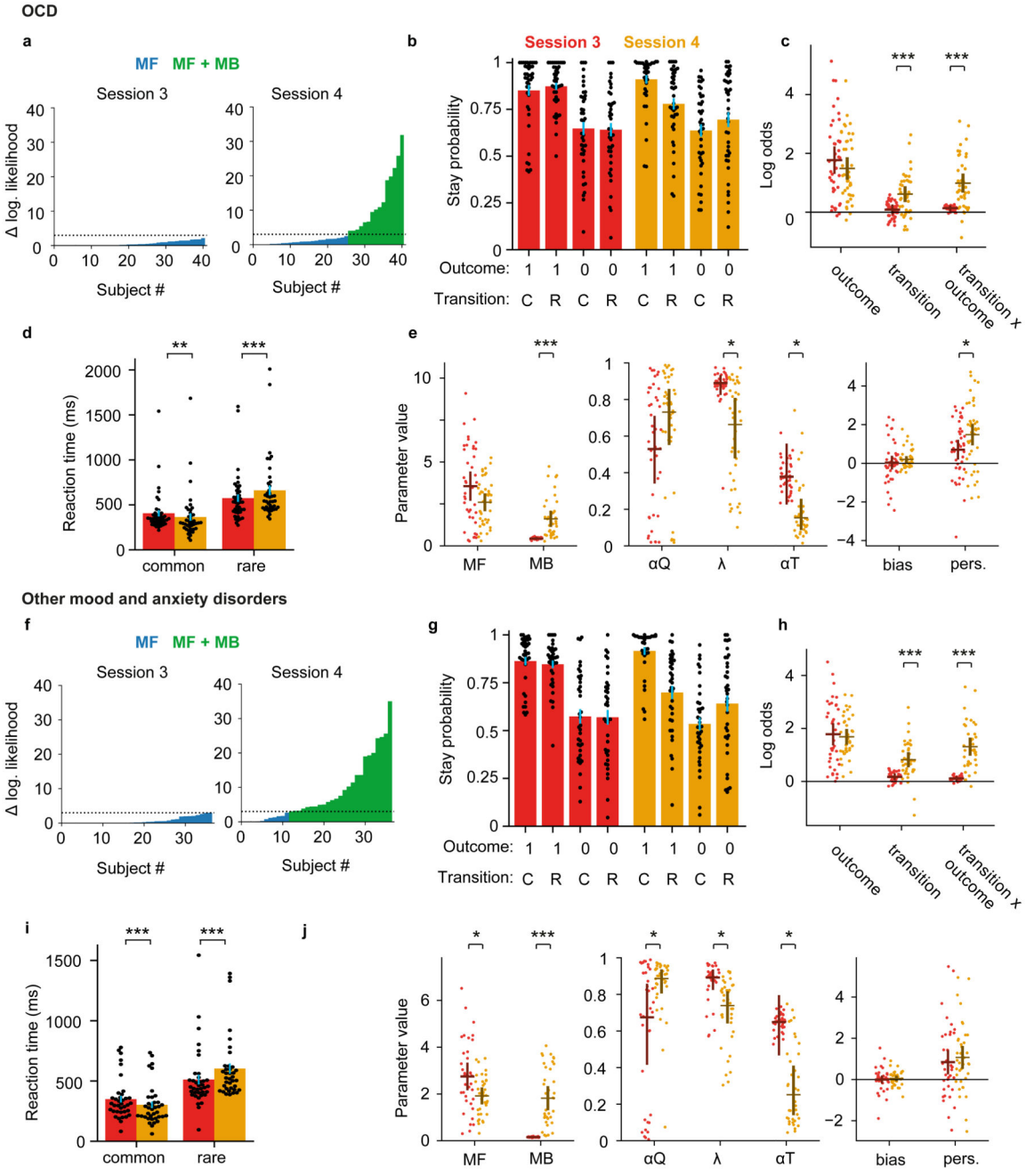

Figure 5. Explicit knowledge increases model-based control in OCD.

Effects of debriefing in 41 individuals with OCD (panels **a-e**), and 37 individuals with mood and anxiety disorders (panels **f-j**), who were model-free at session 3 as assessed by likelihood ratio test. **(a, f)** Per-subject likelihood ratio test for use of model-based strategy at session 3 (left panel) and session 4 (right panel), as figure 4a. **(b, g)** Stay probability analysis, as figure 4b. **(c, h)** Logistic regression analysis of stay probabilities, as figure 4c. In the OCD group, debriefing increased the influence of both the transition (null 95% CI [-0.35,0.37], coefficient change=0.56, P<0.001, permutation test) and transition-outcome

Europe PMC Funders Author Manuscripts

interaction (null 95% CI [-0.50,0.51], parameter change=0.88, P<0.001). In the group with mood and anxiety disorders, debriefing increased the influence of both transition (null 95% CI [-0.40,0.41], coefficient change=0.64, P<0.001, permutation tests) and transition-outcome interaction (null 95% CI [-0.57,0.59], coefficient change=1.20, P<0.001). **(d, i)** Second-step reaction times following common and rare transitions. In both the group with OCD and those with mood and anxiety disorders, debriefing increased differences in second-step reaction times between common and rare transition trials (session-transition interaction, OCD group $F_{1,40}$=30.8, P<0.0001, η_p^2=0.43, mood and anxiety group $F_{1,36}$=26.2, P<0.0001, η_p^2=0.42, repeated measures ANOVA **(e, j)** Comparison of mixture model fits, as figure 4e. In the OCD group, following debriefing the influence of model-based action values on choice increased (null 95% CI [-0.63,0.64], parameter change=1.22, P<0.001), the eligibility parameter decreased (null 95% CI [-0.19,0.19], parameter change=-0.24, *P*=0.017), the transition learning rate decreased (null 95% CI [-0.21, 0.21], parameter change=-0.24, *P*=0.019) and the perseveration parameter increased (null 95% CI [-0.67,0.66], parameter change=0.77, *P*=0.023). In the individuals with mood and anxiety disorders, following debriefing the influence of model-based action values on choice increased (null 95% CI [-0.81,0.81], parameter change=1.63, P<0.001), the influence of model-free action values decreased (null 95% CI [-0.69,0.71], parameter change=-0.82, P=0.019), the value learning rate increased (null 95% CI [-0.17,0.18], parameter change=0.21, P=0.015), the eligibility parameter decreased (null 95% CI [-0.15,0.15], parameter change=-0.15, P=0.043), the transition learning rate decreased (null 95% CI [-0.34,0.33], parameter change=-0.38, P=0.024).

Europe PMC Funders Author Manuscripts

Table 1
Sociodemographic and psychometric characterization of study samples

	HV (n=109)	OCD (n=46)	MA (n=49)
Sex (% males)	33%	41%	31%
Age (years)	30.4 (7.1)	34.1 (12.4)	32.6 (13.1)
Education (years completed)	16.2 (2.5)	15.1 (2.9)	14.5 (4.1)
YBOCS total score	1.5 (3.5)	23 (6.4)	2.7 (4.8)
Y-BOCS obsessions	0.6 (1.8)	11 (3.3)	2.1 (3.7)
Y-BOCS compulsions	0.9 (1.9)	12 (3.5)	0.6 (1.9)
STAI-state score	31.5 (8.1)	47.6 (15.4)	47.9 (11.3)
STAI-trait score	30.8 (8)	56.6 (12)	53.1 (10.1)
BDI-II score[a]	4 (4.8)	21.1 (16.2)	24.8 (12.1)
DASS depression score[b]	1.5 (1.8)	7.8 (5.6)	7.9 (4.4)
DASS anxiety score[b]	0.6 (1.2)	5.2 (4.4)	5.8 (4.3)
DASS stress score[b]	2.5 (2.2)	10.4 (4.7)	8.5 (4.5)
Corsi block tapping test - total span[a]	16 (3.1)	15.4 (3.9)	13.1 (2.5)
No-Go errors in Go/No-Go task (n)	11.2 (7.4)	16 (12.5)	21.2 (11.6)
Reaction time in Go/No-Go task (ms)	470 (44.4)	517 (55.1)	510 (55.1)

HV = Healthy volunteers; OCD = Obsessive-compulsive disorder; MA = Mood and anxiety disorders.

[a] only in Lisbon groups

[b] only in New York groups; YBOCS-II = Yale-Brown Obsessive-Compulsive Scale-II; STAI = State-Trait Anxiety Inventory; BDI-II = Beck Depression Inventory; DASS = Depression Anxiety Stress Scales. Data are presented as mean (standard deviation).