natureresearch

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Explicit knowledge of task structure is a primary determinant of human model-based action
Corresponding author name(s): Albino J. Oliveira-Maia

Reviewer Comments & Decisions:

Decision Letter, initial version:

26th October 2020

Dear Professor Oliveira-Maia,

Thank you once again for your manuscript, entitled "Explicit knowledge of task structure is the primary determinant of human model-based action", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

The chief editorial concern is the degree to which the data and results support your conclusions, an aspect that all referees call into question, voicing concerns about the task, the choice of which parts of the task that were analyzed, the appropriateness of the comparisons, and the lack of important support analyses. We would only be able to consider a revision of your work, if these points were addressed in full. This will require not only further analyses, but also collection of additional empirical data.

Second, the reviewers highlight in a number of instances that previous literature is not sufficiently well incorporated, neither in the description of the motivation for and conclusions of the work, nor in the way in which it should inform the analyses.

We appreciate that addressing these issues, in particular in light of the requirement for further empirical support, is a substantive task. We would therefore understand it if you chose to seek publication elsewhere. In that case, please do inform us of your decision.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. I have also attached a template manuscript file that exemplifies our policies and formatting requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. We understand that the COVID-19 pandemic is causing significant disruptions which may prevent you from carrying out the additional work required for resubmission of your manuscript within this timeframe. If you are unable to submit your revised manuscript within 6 months, please let us know. We will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

Marike Schiffer

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: computational cognitive neuroscience, RL, decision-making

Reviewer #2: computational cognitive neuroscience, RL, decision-making

Reviewer #3: computational cognitive neuroscience, RL, decision-making

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
The study by Castro-Rodrigues, Akam et al. looks at the contribution of model-based and model-free control to action selection in a simplified version of the classic 'two-step' task. In particular, they examine whether subjects' behaviour is initially model-based or model-free when subjects receive minimal instruction on the structure of the task. This is potentially of interest because most previous studies have given explicit instruction about the structure of the task, and a recent report from Todd Hare's lab has found that when these instructions are sufficiently clear then humans tend to be exclusively model-based.

They find that most subjects show behaviour that is initially model-free, not model-based. This behaviour persists over several days of repeated exposure to the task, even though subjects show implicit measures of learning the task structure. When subjects are explicitly instructed about the structure of the task, their behaviour becomes predominantly model-based. Their two central conclusions from this are: (i) explicit task structural knowledge determines human use of model-based RL, and (ii) this is most readily acquired from instruction rather than experience.

In addition to this, the authors explore variation of these behaviours between patients with OCD and other disorders. While there are some differences in the degree to which OCD patients use model-based control, in that uninstructed behaviour is even more biased towards model-free behaviour in these patients, the overall pattern of behaviour is similar to healthy controls (in other words, the patients become more model-based with instruction).

The paper overall provides an interesting contribution to the study of this task, and in some ways complements the recent study by Silva and Hare, by showing what behaviour is like in the "naïve" state of not knowing anything about the task structure. Overall, I thought that the data were well analysed/modelled and clearly presented.

On the other hand, I was left slightly concerned about whether the two central conlusions mentioned above can really be generalised to all learning tasks where the structure of the task is unknown, or whether the conclusions are actually far more narrow and only apply to the situation studied here. I outline these concerns below.

1. My first concern is that subjects performed this task exceedingly quickly, without much deliberation at all (reaction times were typically below 400ms, and subjects completed 1200 trials per session). This leads to a concern that participants were optimising speed over accuracy in completing the task – in which case, a model-free controller might well be the optimal strategy. The task appears to be entirely self-paced (although I couldn't find details on the duration of reward presentation, duration of inter-trial interval, duration of interval between light 1 being extinguished and light 2 appearing, etc. in the methods), and subjects may well not be trying to optimise number of rewards but instead trying to complete the task reasonably quickly. One possible way to address this would actually be to collect a small additional dataset where the number of trials is brought down considerably (e.g. 300 trials per session), and the pace of the task is slowed right down (e.g. each light presentation event lasts 1 second before the subject can respond, and there is a 1 second ITI). In such circumstances, would subjects still appear entirely model-free until they were instructed otherwise? My suspicion is that they would not. Although I accept that the timings of the task are unchanged in session 4 (the instructed session), the instructions are essentially telling participants that they need to pay attention to these features when performing the task (and indeed, they become more deliberative on rare transitions in these circumstances (figure 4d).
2. More broadly, it is unclear whether the general conclusion that participants need explicit task knowledge in order to use model-based RL is really true – or whether it is only true in a task with a very limited state space, like the present task. Model-based control seems far more important in tasks with much larger state spaces than the one considered here. So while the present study provides interesting insights into this particular task, and provides an interesting comparison with the many previous studies that have used this task, it is unclear whether it can be used to draw a broader conclusion about the use of model-based versus model-free control per se.
3. Figure S3 shows that in about 10% of trials, participants make invalid presses at the second stage. It would be interesting to know if these invalid presses reflect implicit learning of the transition probabilities (i.e. whether they are more common on rare transitions than common triansitions), and if so, when this develops across the different sessions (and whether it is at all predictive of the degree of model-based control, or the strength of the common/rare reaction time effect in figure 2e). The fact that subjects appear to be learning the model structure implicitly makes me feel more strongly that if participants were forced to become slower in their responses (comment 1, above), they would likely become more model-based.
4. In a version with changing probabilities, which was run as a pilot study, the authors conclude that subjects were not really able to learn anything at all about the task structure. But a comparison of supplementary figure S6a and figure S6f suggests that a small subset of subjects could learn the task structure, and the proportion of participants didn't really differ between instructed vs. non-instructed versions.

Reviewer #2:
Remarks to the Author:

In this manuscript, Castro-Rodrigues and colleagues aim to study the effect of extensive task instruction (or the lack of it) on one of the most commonly used experimental paradigm related to model-based vs. model-free learning in healthy individuals as well as individuals with OCD and other mood and anxiety disorders.
The authors use a simpler version of the two-step learning task by Daw et al 2011, by presenting all choice options with four circles at once and removing the second-step choice while keeping the main components of the task in terms of state transitions.
The main finding is that in the absence of explicit instruction about the structure of the task, normal and clinical groups predominately adopt a model-free approach while a minority of subjects slowly moved toward model-based approach over time. Moreover, they only found a small difference between healthy and clinical groups in terms of the learning strategy.

It is refreshing to see that finally there is an investigation into the effect of explicit/extensive task instructions in the original version of a task that has caused a lot of controversy about model-based vs. model-free learning, and how much of previous results were dependent on such task instructions. I think this is an important study which is executed and analysed well, and has important implications for studying learning in healthy and clinical populations. The manuscript is clearly written and the main claims of the study are supported by the results presented.

I mainly have a few clarifying questions about the methods/results, some of the side claims and broader implications of the study, and finally some suggestions for additional analyses discussions.

-- I wonder why the authors did not use the original task of Daw et al. I could be wrong but I assume because it is almost impossible to perform the original task without explicit/extensive instructions. If so, what we have learned from many studies using the previous task? There is an interesting paper by Collins and Cockburn (Nat Rev Neuroscience 2020) that is worth discussing.

-- Authors mention that the Changing version of the task was too complex and that is why they focused on the Fixed version in most of their analyses (although they provide results based on the Changing version as well).
Considering that blocks would end in Fixed version of the task, could not participants use this signal to learn faster?
For example, did the time to reach criterion become shorter over block of trials?

Also considering this task structure would not make more sense to analyse data across blocks instead of arbitrary sessions of 300 trials?
Authors show one aspect of behaviour over time (in Figure S1) but not the main analyses (e.g. probability of stay, RT, etc.). I think including such analyses can be more informative, specially about transition from MF to MB.

-- There are a few generally untested ideas about the utility of MB and MF RL, which appears in the introduction of this manuscript as well (lines 44-49).
Is it true that model-based allows behavioural flexibility? If this is the case, MB learning should be higher in the changing environment compared to the fixed version case because more flexibility is needed in the former; Is there any evidence this true?

I addition, authors mention that model-free learning allows rapid action selection but uses information less efficiently. Is there any support for this in the current experiment? If anything, in MF learning both action values can be updated on each trial whereas in MB learning only value related to the observed transition can be updated, making MF faster and more efficiently.
All these claims can be tested in the current study by comparing learning rates and weight of MB vs, MF learning between the Fixed and Changing versions.

-- One of the main points of the current study is that MF learning is more prominent and adopted first. Two recent studies using similarly complex task with multidimensional stimuli (and with no explicit instructions) have shown that feature-based learning, which resembles model-based is the starting learning strategy before transitioning to more accurate object-based learning (Farashahi et al, Nat Comm 2017, Cognition 2020). How do authors see the results of these studies reconcile with their findings?

-- The main results are quite similar across healthy subjects and individual with OCD and other mood disorders (with some minor differences.
Does this mean that this task (and the original task) are not very useful to study changes in MB vs. MF learning in the clinical population, or this dichotomy is not the best way to look at reinforcement learning (see Collins and Cockburn, Nat Rev Neuroscience 2020)?

-- Was there any significance difference in the learning rates between fixed version and changing version of the task. I ask this because there is an intuition that volatility should increase learning rates, and although a study by Behrens et al (Nat Neuroscience 2007) has provided evidence for such an intuition Farashahi et al, Nat Hum Behaviour 2019 found no evidence for it.

-- Was there any significance difference in the learning rates between control and clinical groups?

-- In some places, authors rely in 0.05 as the threshold for statistical significance. Considering the number of comparisons (e.g., between model parameters) I don't think 0.05 is an appropriate threshold.

-- Why there is no comparison between stay probability in different conditions (to prove the unbalanced pattern of stay probability as a function of outcome and transition)?

-- It seems to be a general increase in stay probability after debriefing as a side effect. What could be a reason for this?

-- Authors focus on participants who did not use MB learning in session 3 (e.g. in Figure 4), but what did happen to participants who adopted MB after they received instruction?

Minor

-- Figure 4 caption: "in control"

-- Figure 5 caption: "in OCD and other mood disorders?"

Reviewer #3:
Remarks to the Author:
In the present manuscript, Castro-Rodrigues and colleagues used a modified version of the ‘two-step task’ task to assess contributions of model-based (MB) versus model-free (MF) systems to reinforcement learning in humans. Specifically, they tested to what degree humans deploy MB vs MF learning, first in the absence of any explicit instruction about the task structure, and then after fully explicit debriefing. Both healthy humans and individuals suffering from OCD (and another clinical control sample) were tested. The key finding according to the authors is that uninstructed behaviour was model-free, with model-based control only emerging over time in a subset of participants. The latter was less pronounced in the OCD group. All groups showed stronger model-based behaviour after receiving explicit debriefing on underlying task structure.

The manuscript is well written and addresses an important and interesting question, particularly given influential earlier reports that increased MF control may be a common feature of compulsive behavioural disorders (Voon et al., 2014). However, I am concerned whether the conclusions drawn by the authors are justified by the data and the study design.

1) One of the main findings is that uninstructed behaviour was dominantly model-free. This is a strong statement given the results and the analyses presented. In my view, what can be said with confidence is that subjects did not use the *true* model of the task - and this is the only test of 'model-based behaviour' the authors are presenting. Subjects could have used a completely different model of the task which would not be evident from testing to what degree they used the optimal model. Or indeed volunteers could have tried out different models of the task throughout the three sessions. For instance, as you present in Fig. S1, quite a proportion of people keep pressing invalid keys for a considerable period throughout at least the first session. This implies that it took subjects fairly long to understand even the most fundamental task characteristics. The authors write about there being no evidence of participants using task structure early on (line 337/338) - but given the above, everything else would be highly surprising!

Indeed, da Silva and Hare (which is also cited here) have shown that providing subjects with inaccurate models of the task evokes MB behaviour that can appear completely model-free. They also describe how behaviour which appears to be a hybrid of model-based and model-free could also be explained by a set of different algorithms (see also Collins & Cockburn, 2020). In this context, it would be important to see model comparisons that include competing models as well as model simulations on the eventual winning model.

2) An alternative model employed by participants could be that, rather than assuming a probabilistic, but fixed mapping of first-step choices to second-step states, participants might assume this mapping to be fully deterministic, but highly volatile. Under such a model, following a rare transition,

participants would infer that the underlying latent state has changed and assume that their first-step choice would now lead them to the state 2 they just observed. An agent using such a model would appear exactly like a completely MF agent. This could also be a potential explanation for the effects seen in the OCD group in figure 3 (A and D). It shows that, in OCD, there is an increase in MF control (Gmf) from session 1 to session 3 - and likewise an increase in the effect of outcomes. Again, I am not convinced that such a pattern is fully indicative of MF behaviour. As discussed above, the increase in P(stay) following rewards (irrespective of transition type) could as well be obtained from an inaccurate world model. Thus, such a pattern may actually indicate *increased* MB control in OCD!

3) I appreciate the motivation to simplify the original two-step task. However, by removing the choice alternatives at the second-step state, the task, at least to subjects, may lose its key characteristics as a multistep decision problem, but instead be perceived as a one-stage decision problem with probabilistic rewards (see my comment above). This might have influenced the perceived importance of forming a model and the consideration of the sequential structure of the task. Additionally, as the authors also note, the new design greatly reduces working memory load. Relationships between working memory capacity and task complexity on one hand and decision strategy on the other have been reported previously (Otto et al., 2013; Kim et al., 2019), which may also contribute to the apparent absence of MB behaviour.

4) Furthermore, it was not clear to me why the authors opted for a fixed spatial-motor mapping instead of presenting two distinct visual stimuli with random mapping to top/bottom position on the screen. As the authors acknowledge in the discussion, such a fixed mapping of spatial position to effector may have further encouraged the use of a habitual/model-free/S-R strategy.

5) Modelling:
a) it would be re-assuring to see whether the fitted parameters can be recovered from simulated data generated using the fitted parameters (parameter recovery).
b) Unlike in Daw 2011, there are separate weights (Gmb, Gmf) for the contributions of the MB and MF systems, respectively (rather than Gmf = 1 – Gmb). This seems plausible to me, as there may well be participants in which both MB and MF is low (e.g. purely stochastic or perseverative choice), but since this is also a - minor - departure from the original analyses, it would be helpful to briefly justify this in the text.
c) In figure 2F, the modelling results do not support the logistic regression - there is no difference in Gmf and Gmb between session 1 and 3.
d) Bias parameter in the model: is my understanding correct that this indicates an action bias for the top circle? If so, why is it B = 1 for the high and B = 0 for the low action? Would this not mean that the model can have a bias for the upper circle, or no bias at all? Shouldn't B = −1 for the low option to also allow for a bias toward the lower option?
e) Decreases in the eligibility trace parameter lambda correlate, across subjects, with increase in MB control - is this naturally arising because as lambda --> 0, there is no more update in the MF system? In other words, is it possible, for a given subject, that there are two (probably very similar) local optima in the parameter space, at low (high) values for lambda and high (low) values for Gmb?

6) In figure 2F, the modelling results indicate no difference between MB and MF contributions to behaviour. This is at odds with the logistic regression results presented in 2D (and the p(stay) in 2C) which indicates a clear transition X outcome interaction. This again raises the issue to what extent the model with the fitted parameters is able to recapitulate the actual pattern in subjects' choice behaviour (related to my question above regarding parameter recovery)

7) In the analyses investigating the effects of providing the full task structure, it seems that out of the 57 healthy individuals not showing MB behaviour yet, n = 41 were assigned to the debriefing group whereas only n = 16 were not debriefed. Why were they evenly allocated to both groups? Comparing the proportion of subjects acquiring MB control appears odd given the highly unbalanced group size?

8) Why are people diagnosed with a wide variety of mood and anxiety disorders put together as one group? These disorders are characterized by several different symptoms which could also be expected to influence MB/MF control of behaviour in different ways. I guess it also does not serve to “investigate potential contributions of medication or unspecified mood and anxiety symptoms” for the same reasons. Further details on psychotropic medication would be needed to evaluate their comparability with respect to medication.

9) RT at 2nd stage are faster in common compared to rare transitions, and this effect is present already in session 1 (but seems to become larger in session 3). Is this an indication that subjects have learnt the state transition probabilities in session 1 already?

10) Before debriefing participants, did you question them about what kind of strategy they used? While this would probably hard to quantifiy, it might at least give some qualitative hints as to what beliefs they had acquired (and what models of state space they applied) during the 900 uninstructed trials.

11) It is unclear to me why the neutral blocks existed? What was their purpose?

Minor:
1) In the introduction (Line 79 onwards) you write: "However, no studies have explored behaviour on multi-step tasks in the absence of information about task structure, in either healthy or clinical populations" <-- I can think of one study (Gläscher et al. 2010, Neuron) that used Tolman-style latent learning in an abstract multi-step T-maze. I think this is also without explicit instruction and likewise looks at the contribution of MF and MB systems over time.

2) In the discussion (#341): "When learning from experience, individuals with OCD were impaired in their use of model-based control and biased towards a model-free strategy." This makes it sound as if this was only true for OCD, but it likewise applies for healthy controls.

3) Line 283/285: "Increased use of model-based RL after debriefing was confirmed by model fitting (Figure 5d), which showed increased influence of model-based action values on choice" <-- I guess this should be Figure 5e?

Author Rebuttal to Initial comments

Reviewer #1

Remarks to the Author:

The study by Castro-Rodrigues, Akam et al. looks at the contribution of model-based and model-free control to action selection in a simplified version of the classic 'two-step' task. In particular, they examine whether subjects' behaviour is initially model-based or model-free when subjects receive minimal instruction on the structure of the task. This is potentially of interest because most previous studies have given explicit instruction about the structure of the task, and a recent report from Todd Hare's lab has found that when these instructions are sufficiently clear then humans tend to be exclusively model-based.

They find that most subjects show behaviour that is initially model-free, not model-based. This behaviour persists over several days of repeated exposure to the task, even though subjects show implicit measures of learning the task structure. When subjects are explicitly instructed about the structure of the task, their behaviour becomes predominantly model-based. Their two central conclusions from this are: (i) explicit task structural knowledge determines human use of model-based RL, and (ii) this is most readily acquired from instruction rather than experience.

In addition to this, the authors explore variation of these behaviours between patients with OCD and other disorders. While there are some differences in the degree to which OCD patients use model-based control, in that uninstructed behaviour is even more biased towards model-free behaviour in these patients, the overall pattern of behaviour is

similar to healthy controls (in other words, the patients become more model-based with instruction).

The paper overall provides an interesting contribution to the study of this task, and in some ways complements the recent study by Silva and Hare, by showing what behaviour is like in the "naïve" state of not knowing anything about the task structure. Overall, I thought that the data were well analysed/modelled and clearly presented.

We thank the reviewer for the positive assessment of our manuscript.

On the other hand, I was left slightly concerned about whether the two central conclusions mentioned above can really be generalised to all learning tasks where the structure of the task is unknown, or whether the conclusions are actually far more narrow and only apply to the situation studied here. I outline these concerns below.

1. My first concern is that subjects performed this task exceedingly quickly, without much deliberation at all (reaction times were typically below 400ms, and subjects completed 1200 trials per session). This leads to a concern that participants were optimising speed over accuracy in completing the task – in which case, a model-free controller might well be the optimal strategy. The task appears to be entirely self-paced (although I couldn't find details on the duration of reward presentation, duration of inter-trial interval, duration of interval between light 1 being extinguished and light 2 appearing, etc. in the methods), and subjects may well not be trying to optimise number of rewards but instead trying to complete the task reasonably quickly. One possible way to address this would actually be to collect a small additional dataset where the number of trials is brought down considerably (e.g. 300 trials per session), and the pace of the task is slowed right down (e.g. each light presentation event lasts 1 second before the subject can respond, and there is a 1 second ITI). In such circumstances, would subjects still appear entirely model-free until they were instructed otherwise? My suspicion is that they would not. Although I accept that the timings of the task are unchanged in session 4 (the instructed session), the instructions are essentially telling participants that they need to pay attention to these features when performing the task (and indeed, they become more deliberative on rare transitions in these circumstances (figure 4d).

We agree with the reviewer that, in principle, due to the self-paced nature of the original task, and overall fast reaction times exhibited by subjects, they might have been optimising speed over accuracy and hence opting for a model-free strategy. We directly addressed this in the revised manuscript by following the reviewer's suggestion and gathering an additional dataset in which we slowed down the pace of the task by introducing a 1 second delay between options being highlighted and being active for selection (cued by a colour change from pale to bright yellow), at both the first and second step, in addition to a 1 second inter-trial interval. In this dataset, 20 newly recruited healthy volunteers completed a total of 600 trials each, split into 3 sessions of 150 trials prior to debriefing, and an additional session of 150 trials after debriefing (the original task used 4 sessions each of 300 trials). Despite the slower pace and hence more time for deliberation, the new data strikingly recapitulated our original findings, including:

- A large influence of trial outcome but no effect of transition-outcome interaction at the first session, consistent with a model-free strategy (Figure S4, reproduced below).

- A small increase in transition-outcome interaction by session 3, consistent with increased use of model-based RL with experience (Figure S4b). As in the original task, this appears to be driven by a minority of subjects, as a likelihood ratio test on session 3 data supported a mixed model-based + model-free strategy over a model-free only strategy in only 3 among the 20 subjects.

- A large increase in the use of model-based RL after receiving explicit information about task structure. As in the original task, this was accompanied by a decrease in the RL model eligibility trace parameter, which we think reflects changes to subjects' representation of the task-state space affecting model-free value updates (Figure S5, reproduced below).

Overall these data confirm our main results that, in unfamiliar domains, subjects initially rely on a model-free strategy, that they are surprisingly slow to learn to use a task model from experience, but learn much more readily from explicit description.

These data are detailed in the revised manuscript as:

Results 193-208: *A possible reason why model-free control might predominate is that subjects 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). Subjects 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 (P < 0.001, permutation test) in session 1, but no effect of transition (P=0.23) or transition-outcome interaction (P=0.92) (Figure S4a,b). Also similarly to the self-paced task, the effect of transition-outcome interaction on stay* *probability increased between session 1 and 3 (P=0.008), as assessed by the logistic regression (Figure S4b), 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 (Figure S4d), 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 subjects.*

Results 258-260: *Similar effects of debriefing on the transition-outcome interaction were found among the 17 healthy volunteers performing the slow-paced version of the task that were not using model-based RL significantly in session 3 (Figure S5a-c).*

Supplementary figure S4. Learning effects in slow-paced task. a) Stay probability analysis. **b)** Logistic regression analysis of stay probabilities. **c)** Reaction times after common and rare transitions in session 1 and 3. **d)** Comparison of mixture model fits between session 1 and session 3. RL model parameters: MF, Model-free strength; MB, Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.

Supplementary figure S5. Effects of explicit knowledge in the slow paced task. **a)** Per-subject likelihood ratio test for use of model-based strategy on session 3 (left panel) and session 4 (right panel). **b)** Stay probability analysis. **c)** Logistic regression analysis of stay probabilities. **d)** Reaction times after common and rare transitions in session 1 and 3. **e)** Comparison of mixture model fits between session 1 and session 3. RL model parameters: MF, Model-free strength; MB, Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.

2. More broadly, it is unclear whether the general conclusion that participants need explicit task knowledge in order to use model-based RL is really true – or whether it is only true in a task with a very limited state space, like the present task. Model-based control seems far more important in tasks with much larger state spaces than the one considered here. So while the present study provides interesting insights into this particular task, and provides an interesting comparison with the many previous studies that have used this task, it is unclear whether it can be used to draw a broader conclusion about the use of model-based versus model-free control per se.

We agree that it is possible that the size of the task state space might affect use of model-based RL. It is worth noting that the task did incentivise the use of model-based RL, as shown by the positive correlation we observed between measures of model-based choice and total rewards obtained, demonstrating that use of model-based RL is important in our task. We also note that, while larger state spaces might increase payoffs for using model-based RL, they also make both model-learning and planning substantially more difficult, so it is not obvious that model-based RL would necessarily be utilised more or earlier in larger state spaces. Nonetheless we agree it is important to be transparent about this limitation of our study, which we now discuss in the revised manuscript file as:

Discussion, lines 528-532: *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.*

We have also softened the language in the title and abstract to state that explicit knowledge is 'a' rather than 'the' primary determinant of human model-based action.

3. Figure S3 shows that in about 10% of trials, participants make invalid presses at the second stage. It would be interesting to know if these invalid presses reflect implicit learning of the transition probabilities (i.e. whether they are more common on rare transitions than common triansitions), and if so, when this develops across the different sessions (and whether it is at all predictive of the degree of model-based control, or the strength of the common/rare reaction time effect in figure 2e). The fact that subjects appear to be learning the model structure implicitly makes me feel more strongly that if participants were forced to become slower in their responses (comment 1, above), they would likely become more model-based.

We thank the reviewer for raising this important question, addressed via a new analysis, added to figure S1, comparing the rate of invalid presses at the second-step (left or right actions), following common and rare transitions. In the first session the rate of invalid presses following common and rare transitions were not significantly different, but by the third session subjects made significantly more errors following rare transitions. This provides further evidence for implicit learning of task structure, in addition to the previously reported reaction time differences following common vs rare transitions, which are significant even in the first session.

We have also added an additional supplementary figure (Figure S3, reproduced below) looking at correlations, across subjects, between measures of model-based choice (model-based weight from the RL model and transition-outcome interaction from the regression) and measures of implicit learning of the transition structure (differences in second-step reaction times and rates of invalid responses following common vs rare transitions). Interestingly, choice-based and implicit measures of task structure learning, reaction-time in particular, were significantly correlated at all time-points evaluated (sessions 1,3 and 4). This appears to be driven, at least in sessions 1 and 3, when subjects are learning only from experience, by the handful of subjects whose choices are more model-based.

Taken together, we think these data support a partial dissociation between task structure learning at the motor and cognitive level, with motor-level learning occurring faster and more consistently across the population, as evidenced by the robust reaction time effects at session 1 when choices are model-free, but with some interaction between the levels evidenced by the cross-subject correlations.

We agree with the reviewer that evidence for implicit task structure learning further motivates the control experiment we performed in response to comment 1 above, in which we forced subjects to do the task at a slower pace.   As discussed above, choice behaviour in this slow-paced task was similar to that in our original data. We think these data are consistent with a dissociation between task structure learning at motor and cognitive levels, in that robust implicit learning of task structure, as evidenced by motor performance, does not necessarily imply that cognitive systems have a usable task model.

We discuss these data in the revised manuscript as:

Results lines 143-147: *Although state transitions in session 1 did not influence subsequent first-step choices, 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; t66=7.81, P<0.0001, 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.*

Results lines 184-192: *Key-press reaction times at the second-step became faster overall between session 1 and 3 (main effect of session P<0.0001, repeated measures ANOVA), but this was more pronounced following common than rare transitions (session-transition interaction P=0.008; Figure 2e). Additionally, by session 3 the rate of invalid key presses was significantly higher following rare than common transitions (P=0.002, Wilcoxon signed rank test, Figure S1). Therefore, both choice-based and implicit measures showed evidence of learning about the transition structure between session 1 and 3. Model-based choice was significantly correlated across subjects with the rare-common reaction time difference at both session 1 & 3 (P<0.001, figure S3), suggesting interaction between learning at motor and cognitive-levels, thought this appeared to be driven by the minority of subjects whose choices were more model-based.*

Discussion, lines 482-489: *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[56]. Intriguingly, implicit and choice-based measures of task structure learning were correlated across subjects, even at session 1, suggesting that this dissociation is only partial, with cognitive task models potentially informed by earlier motor-level learning.*

Supplementary figure S3. Correlation between choice and implicit measures of task structure learning. Correlation across subjects between different measures of task structure learning in the fixed task at sessions 1, 3 and 4. Two choice-based measures were used; the model-based weight parameter from the RL model fit and the transition x outcome predictor loading from the logistic regression, and two implicit measures; the difference in second-step reaction times following common vs rare transitions, and the difference in the rate of invalid second-step responses (e.g. pressing left when the right state was active) following common vs rare transitions. Points show individual subjects. Lines show linear fit with 95% confidence interval on fit indicated by shaded region. In both session 1 and 3, when subjects are learning only from experience, both measures of model-based choice are correlated with the rare-common reaction time difference.

4. In a version with changing probabilities, which was run as a pilot study, the authors conclude that subjects were not really able to learn anything at all about the task structure. But a comparison of supplementary figure S6a and figure S6f suggests that a small subset of subjects could learn the task structure, and the proportion of participants didn't really differ between instructed vs. non-instructed versions.

The reviewer is correct that some individual subjects on the changing task did show evidence of model-based RL in the last session (session 4), as assessed by the likelihood ratio test. The reviewer is also correct that informing subjects about the structure of the changing task did not significantly increase the use of model-based RL, since the proportion of participants using model-based RL on session 4 was the same independent of debriefing (2/12 vs. 4/24). Also, the number of participants that started using model-based RL on session 4 (6/36) was similar to those using model-based RL on session 3 (6/42). The lack of effects of debriefing presumably reflects the fact that participants did not understand the debriefing, or felt the task was too complex to make the effort. This was not the case in the fixed task, where debriefing dramatically boosted use of model based RL. This was one of the reasons why we choose the fixed task for the majority of our experiments. The other reason was that, taking the population performing the changing task as a whole, we saw no evidence of a significant increase in the use of model-based RL during the uninstructed sessions, i.e. we did not see the increase in transition-outcome interaction that we saw in the fixed task. We have clarified this in the revised manuscript as:

Results, lines 118-121: *The Changing version proved too complex for most subjects, 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 Supplementary information for details), so we subsequently focused on the Fixed task, which is used for all data and figures in the main text.*

Supplementary information, lines 11-12*: The proportion of subjects for whom a likelihood ratio test indicated model-based RL was being used in session 3 (6/42) was similar to that observed in the Fixed task (10/67).*

Supplementary information, lines 25-29: *These data suggest that, while model-free RL was dominant for the non-instructed sessions of both tasks, the dynamically changing action-state transition probabilities in the Changing task further reduced the ability to learn a model-based strategy. This is consistent with uncertainty based arbitration between model-based and model-free control[33,4], as the changing transition probabilities would be expected to increase uncertainty in the model-based system.*

Reviewer #2:

Remarks to the Author:

In this manuscript, Castro-Rodrigues and colleagues aim to study the effect of extensive task instruction (or the lack of it) on one of the most commonly used experimental paradigm related to model-based vs. model-free learning in healthy individuals as well as individuals with OCD and other mood and anxiety disorders.

The authors use a simpler version of the two-step learning task by Daw et al 2011, by presenting all choice options with four circles at once and removing the second-step choice while keeping the main components of the task in terms of state transitions.

The main finding is that in the absence of explicit instruction about the structure of the task, normal and clinical groups predominately adopt a model-free approach while a minority of subjects slowly moved toward model-based approach over time. Moreover, they only found a small difference between healthy and clinical groups in terms of the learning strategy.

It is refreshing to see that finally there is an investigation into the effect of explicit/extensive task instructions in the original version of a task that has caused a lot of controversy about model-based vs. model-free learning, and how much of previous results were dependent on such task instructions. I think this is an important study

which is executed and analysed well, and has important implications for studying learning in healthy and clinical populations. The manuscript is clearly written and the main claims of the study are supported by the results presented.

We thank the reviewer for the positive assessment of the manuscript.

I mainly have a few clarifying questions about the methods/results, some of the side claims and broader implications of the study, and finally some suggestions for additional analyses discussions.

-- I wonder why the authors did not use the original task of Daw et al. I could be wrong but I assume because it is almost impossible to perform the original task without explicit/extensive instructions. If so, what we have learned from many studies using the previous task? There is an interesting paper by Collins and Cockburn (Nat Rev Neuroscience 2020) that is worth discussing.

The reviewer is correct that we did not use the original task because we predicted that it would be too complicated for participants to learn uninstructed. We intended to simplify the task as much as possible while preserving the structure necessary to dissociate model-based and model-free RL, in order to optimize the chance of observing model-learning.

While we think our findings provide important insights into when humans do and do not use model-based RL, we do not think they invalidate findings with the original task. Rather, they help to understand where those finding are likely to apply and where they may not. Specifically, our data speaks to situations where subjects must learn from experience to act in an unfamiliar domain, and show that behaviour in these circumstances is very different to that in situations where subjects are given detailed information about task structure. Both of these situations occur in real-life decision making, and are hence important to understand, though we think the uninstructed situation is under-represented in laboratory studies. Our study contributes to clarifying processes associated with uninstructed learning and the effects of instruction.

We agree the Collins and Cockburn paper is worth discussing and do so as:

Discussion, lines 522-528: *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, 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 subjects 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.*

-- Authors mention that the Changing version of the task was too complex and that is why they focused on the Fixed version in most of their analyses (although they provide results based on the Changing version as well). Considering that blocks would end in Fixed version of the task, could not participants use this signal to learn faster? For example, did the time to reach criterion become shorter over block of trials?

Following the reviewer's suggestion, we quantified block length distributions on the fixed and changing version of the task during sessions 1-3, when subjects are learning from experience, and found that they were very similar (Fixed task 55.7 ± 32.1 trials, Changing task 55.7 ± 32.2 trials, mean ± SD). In these sessions most subjects are using model-free RL, and, in both tasks, model fits to these initial sessions indicate that the model-free update of first-step action values was primarily driven by the trial outcome (rewarded or not) rather than the value of the second-step state (this is indicated by the model's eligibility trace parameter λ being close to 1). For a model-free agent which updates the first step action values only from the trial outcome, irrespective of the second step state reached, the fixed and changing tasks are identical. Independently of whether the reward probabilities or transition probabilities reverse, the update required to the first-step action value is the same.

Importantly, our assessment that the changing version was too complex was strongly influence by the effects of debriefing in both task versions. Specifically, debriefing significantly increased the proportion of individuals developing model-based RL in session 4 for the Fixed task (52.1% for the debriefed group and 6.3% for the non-debriefed group), but not for the Changing task (16.7% both for debriefed and non-debriefed groups). This has been clarified in the revised manuscript at lines 118-121, which read:

The Changing version proved too complex for most subjects, 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 supplementary information for details), so we subsequently focused on the Fixed task, which is used for all data and figures in the main text.

Regarding the reviewers' question about whether subjects got faster at adapting to reversals with task experience, we looked at this by examining the choice probability trajectories following reversals (see figure below) but did not observe significant changes between session 1 and 3 in either task (permutation test for difference in exponential fit P>0.17).

Also considering this task structure would not make more sense to analyse data across blocks instead of arbitrary sessions of 300 trials? Authors show one aspect of behaviour over time (in Figure S1) but not the main analyses (e.g. probability of stay, RT, etc.). I think including such analyses can be more informative, specially about transition from MF to MB.

We thank the reviewer for this suggestion, which we considered very carefully. On balance we think that analysing the data in blocks is probably not straightforward, because we do not see a way to look at the time-course of learning across blocks which does not either a) exclude a large number of subjects, or b) incur major biases due to including different subjects at different time points. This is because the number of blocks completed by subjects for each of the first 3 sessions varies over a wide range (shown below for the fixed task). If we divide the data based on block number, then either we will have many more subjects in the early data, or we will have to exclude a substantial quantity of either trials or subjects to balance the data. This would be further complicated by the fact that the number of completed blocks will be correlated with their behavioural strategy.

We think that grouping the data into early and late learning by session number, and hence trials completed, is a reasonable and principled approach to categorising the stage of learning. The choice of 300 trial lengths for sessions was, as the reviewer correctly points out, arbitrary, but given that this is what we used for the experiment it is a natural choice to also use for the analysis. It is also worth noting that we obtained qualitatively similar findings in the slow paced control experiment in which the length of sessions was 150 trials.

-- There are a few generally untested ideas about the utility of MB and MF RL, which appears in the introduction of this manuscript as well (lines 44-49).

Is it true that model-based allows behavioural flexibility?

Our comments about the utility of model-based and model-free RL were aiming to outline the computational properties of these different classes of algorithm, rather than their behavioural consequences. We apologize if this was not clear, we have modified the text to ensure we correctly cite the relevant literature when we make these statements. Regarding computational properties, we think it is an accurate characterisation to say that model-based RL algorithms can adapt to changes in the environment faster than a model-free RL algorithm. The reason for this is that a local change in the environment can change the optimal policy globally – e.g. if the rewarded goal location in a maze changes, this may require different decisions at locations far removed from the goal. When a model-based algorithm which has learned a model of the environment observes a local change in either the reward or transition structure, it can calculate the consequences of this for policy across the entire environment, albeit at the computational cost of planning. A model-free algorithm by contrast must learn the new correct policy via experience.

If this is the case, MB learning should be higher in the changing environment compared to the fixed version case because more flexibility is needed in the former; Is there any evidence this true?

In accordance with the above discussion regarding block lengths being similar in the two versions of the task, we expect the fixed and changing tasks should require a similar level of behavioural flexibility, given that the correct choice changes equally often in both. The only difference is that in the fixed task this is always because the reward probabilities have changed, whereas in the changing task this can be because the reward probabilities have changed, or because the transition probabilities have changed. It results that the adaptation of model-based control is spread slightly differently between the two tasks. However, as discussed above, from the perspective of a model-free RL algorithm these tasks look very similar, and a majority of subjects are model-free pre-debriefing.

We also note that previous studies have argued for arbitration between model-based and model-free control of behaviour on the basis of the uncertainty in each system about the best policy (Daw et al. 2005 Nature neuroscience 8 (12), 1704-1711, Lee et al. 2014, Neuron 81 (3), 687-699). The non-stationary transition probabilities in the Changing task would be expected to increase uncertainty about the transition probabilities used by the model-based system, which according to this account would be expected to reduce the influence of model-based control on behaviour. Consistent with this, in the Changing task we did not see the increased loading on the transition-outcome interaction predictor with experience that we observed in the Fixed task (Figure 2D, S10B), and removing the model-based component of the RL model had less impact on the fit quality in the Changing task than in the Fixed task (Figure S2A, S10D).

This is now discussed in greater detail as:

Supplementary information, lines 25-29: *These data suggest that, while model-free RL was dominant for the non-instructed sessions of both tasks, the dynamically changing action-state transition probabilities in the Changing task further reduced the ability to learn a model-based strategy. This is consistent with uncertainty based arbitration between model-based and model-free control[33,4], as the changing transition probabilities would be expected to increase uncertainty in the model-based system.*

I addition, authors mention that model-free learning allows rapid action selection but uses information less efficiently. Is there any support for this in the current experiment? If anything, in MF learning both action values can be updated on each trial whereas in MB learning only value related to the observed transition can be updated, making MF faster and more efficiently. All these claims can be tested in the current study by comparing learning rates and weight of MB vs, MF learning between the Fixed and Changing versions.

We appreciate the question regarding efficient use of information with MF vs. MB learning. Indeed, our data is consistent with model-based RL using information more efficiently, since loading on the transition-outcome interaction predictor across sessions 1 to 3 (a measure of model-based RL) was positively correlated with the number of rewards obtained by each subject, for both tasks. This is reported in the main manuscript, and has been revised for clarity, as:

Results, lines 155-158: *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 (rho=0.67, P<0.001), suggesting that subjects who learned a model of the task used information more efficiently and thus obtained rewards at a higher rate.*

This finding is also consistent with current consensus that, in standard model-free RL, only the value of the chosen action is updated at each step, which in the two-step task corresponds to only one of the first-step actions being updated on each trial. By contrast, for a model-based agent, the change in value of the second-step state due to the trial outcome will update the value of both first-step actions according to the subjects' current estimate of how likely they are to transition to that state.

Regarding comparisons between the two tasks, as discussed above, prior to receiving explicit information about task structure, the predominant strategy in both tasks was model-free RL, with the fitted parameters further indicating that trial outcomes (rewarded or not) primarily directly reinforced the preceding first step choice, irrespective of the particular second-step in which the reward was obtained. This is consistent with the observed large main effect of outcome on repeating choice, as observed in the logistic regression. From the perspective of such a 'direct-reinforcement' model-free strategy, which is insensitive to where the rewards are received, the fixed and changing tasks are very similar. Consistent with this we did not see any significant differences in the RL model fits between the fixed and changing task in sessions 1 or 3, and the proportion of participants developing model-based RL at session 3 was similar across the two versions of the task (~14-15%). In any case, as discussed above, the logistic regression and model-comparison analyses provide some evidence for some, although limited, use of model-based RL in the Fixed task, consistent with uncertainty-based arbitration between strategies. Equivalent evidence was not found for the Changing task. Similarities and differences between the two tasks in the uninstructed phase are now discussed in greater detail in the Supplementary information:

Supplementary information, lines 11-12: *The proportion of subjects for whom a likelihood ratio test indicated model-based RL was being used in session 3 (6/42) was similar to that observed in the Fixed task (10/67).*

Supplementary information, lines 25-29: *These data suggest that, while model-free RL was dominant for the non-instructed sessions of both tasks, the dynamically changing action-state transition probabilities in the Changing task further reduced the ability to learn a model-based strategy. This is consistent with uncertainty based arbitration between model-based and model-free control33,4, as the changing transition probabilities would be expected to increase uncertainty in the model-based system.*

-- One of the main points of the current study is that MF learning is more prominent and adopted first. Two recent studies using similarly complex task with multidimensional stimuli (and with no explicit instructions) have shown that feature-based learning, which resembles model-based is the starting learning strategy before transitioning to more accurate object-based learning (Farashahi et al, Nat Comm 2017, Cognition 2020). How do authors see the results of these studies reconcile with their findings?

Feature- vs object-based learning is interesting, and we appreciate the opportunity to discuss our results in the context of the work by Farashahi and colleagues. Our understanding is that the distinction between feature-based and object-based learning is, at least in principle, largely orthogonal to that between model-based and model-free learning. The model-based vs model-free dichotomy concerns what subjects learn to predict about the future. Specifically, whether individuals learn to predict future reward directly (model-free), or to learn to predict future states and likely contingent outcomes, and hence estimate expected future reward indirectly (model-based). The feature- vs object-based distinction by contrast, is to do with whether and how subjects generalise learning among items based on their sensory features. Indeed, one could imagine, for future research, that the task described here could be enriched with stimuli that have multiple feature dimensions, only some of which are relevant. Given our observation that instructions appear to shape subjects' internal representation of which states are important or distinct, even in a task where states were explicitly signalled, it would be very interesting to understand instruction effects in the more ambiguous conditions investigated by Farashahi and colleagues, where relevant features must be inferred.

We discuss this in the revised manuscript as:

Discussion, lines 532-534: *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].

-- The main results are quite similar across healthy subjects and individual with OCD and other mood disorders (with some minor differences.

Does this mean that this task (and the original task) are not very useful to study changes in MB vs. MF learning in the clinical population, or this dichotomy is not the best way to look at reinforcement learning (see Collins and Cockburn, Nat Rev Neuroscience 2020)?

As discussed above response to an earlier question from reviewer 2, we believe that both the original Daw two-step task and the uninstructed version presented here provide valuable and complementary information about human behaviour. They examine distinct situations, both of which are common in real-life decision making. Specifically, our data speaks to situations where subjects must learn from experience to act in an unfamiliar domain, and show that behaviour in these circumstances is very different to that in situations where subjects are given detailed information about task structure. Also as discussed previously, we agree that the limitations pointed out by Collins and Cockburn paper are very much worth discussing and this has been added to the discussion section of the revised manuscript as:

Discussion, lines 522-528: *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, 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 subjects 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.*

We don't think our findings of limited differences between OCD patients and healthy volunteers in their use of model-based RL invalidate the previous findings with the Daw task, among other reasons because the two tasks are distinct, and we did not compare them directly . Rather, we propose that our data suggest that previously unexplored aspects of how tasks are presented can affect the extent and nature of clinical differences, and have proposed potential clinical implications of such findings in the discussion.

-- Was there any significance difference in the learning rates between fixed version and changing version of the task. I ask this because there is an intuition that volatility should increase learning rates, and although a study by Behrens et al (Nat Neuroscience 2007) has provided evidence for such an intuition Farashahi et al, Nat Hum Behaviour 2019 found no evidence for it.

As discussed above, when the subjects are learning from experience, the block lengths in the Fixed and Changing task are very similar, so the volatility of the tasks is matched with regards to the predominantly model-free strategy the subjects employed. We therefore don't think our data speak to the question of whether/how environment volatility affects learning rates. Consistent with this, as we noted in response to an earlier comment from this reviewer, we did not observe significant differences in RL model-fits to the fixed and changing task pre-debriefing.

-- Was there any significance difference in the learning rates between control and clinical groups?

We do not see any significant differences in the effects of either uninstructed experience or debriefing on learning rates between either clinical group and healthy volunteers. Significant effects are reported in the main text and we now report the full set of permutation test results in new supplementary tables S1 and S2 (shown below).

Supplementary table S1 - Differences in learning and debriefing effects between healthy volunteers and individuals with OCD

Model parameters	Learning effects[a]	Debriefing effects[a]
Model-free strength (MF)	0.076	0.96
Model-based strength (MB)	0.40	0.83
Value learning rate (αQ)	0.99	0.42

Eligibility trace (λ)	0.77	0.83
Transition learning rate (αT)	0.88	0.80
Choice bias	0.66	0.20
Choice perseveration	0.48	**0.037**

[a] Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between healthy volunteers (n=67) and individual with OCD (n=46) in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3). P-values for interactions between group and the effect of interest are shown.

Supplementary table S2 - Differences in learning and debriefing effects between healthy volunteers and individuals with mood and anxiety disorders

Model parameters	**Learning effects[a]**	**Debriefing effects[a]**
Model-free strength (MF)	0.65	0.73
Model-based strength (MB)	0.93	0.06
Value learning rate (αQ)	0.49	0.91

Eligibility trace (λ)	0.28	0.72
Transition learning rate (αT)	0.55	0.90
Choice bias	0.44	0.50
Choice perseveration	0.11	**0.001**

[a] Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between healthy volunteers (n=67) and individual with mood and anxiety disorders (n=49) in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3). P-values for interactions between group and the effect of interest are shown.

-- In some places, authors rely in 0.05 as the threshold for statistical significance. Considering the number of comparisons (e.g., between model parameters) I don't think 0.05 is an appropriate threshold.

Our manuscript contains both important positive results and important negative results, i.e. in several places it is the absence of a significant effect that is striking given the prior literature. We therefore think it is important to balance the risk of false positives against that of false negatives, as we do not want readers to come away with the impression that where we did not observe differences where they might be expected, this was because we had a high risk of false negatives due to multiple comparison correction. Our approach in the manuscript has therefore been to report uncorrected P values in full (as opposed to as P<0.05 etc), such that readers can directly evaluate the strength of the statistical evidence. It is worth noting that most of our positive results have very robust P values that would survive multiple comparison correction for multiple model-parameters. We also note that we replicated our key findings concerning behaviour of healthy volunteers to a striking extent in the new dataset gathered with a slow paced version of the task to address reviewer 1’s comments (figure S4 & S5). Specifically, we confirmed that initial behaviour is model-free, that model-based increases in a minority of subjects with experience, and that subsequently providing explicit knowledge strongly boosts use of model-based but also affects model-free value updates.

-- Why there is no comparison between stay probability in different conditions (to prove the unbalanced pattern of stay probability as a function of outcome and transition)?

Given previous work from some of the authors of this paper, we think that reporting statistics based on the logistic regression analysis is the most principled way of quantifying the influence of trial events on the subsequent choice. Specifically, we have used the logistic regression analysis (e.g. Figure 2D) to quantify significant effects of outcome, transition and their interaction on stay probability. This analysis models the same data as shown in the raw stay probability plots, but importantly allows us to incorporate other influences on subjects' choices such as choice biases. Akam et al. (PLOS Comp. Biol. 2015) have previously shown that in two-step tasks with a strong contrast between good and bad options (including the one used here), correlations across trials can cause even model-free agents to show a significant influence of transition-outcome interaction on stay probability, because the difference between chosen and non-chosen action values at the start of the trial is correlated with the subsequent transition-outcome interaction. In that paper it is shown that this can be corrected for by including an additional predictor in the logistic regression analysis of stay probabilities, allowing unbiased estimation of the effects of the trial events on the next trials stay probabilities. We have clarified this approach in the methods section of the revised manuscript as:

Methods, lines 652-655: *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 subjects biases and cross trial correlations (see below) to be taken into account.*

Methods, lines 660-663: *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[42].*

-- It seems to be a general increase in stay probability after debriefing as a side effect. What could be a reason for this?

The reviewer is correct that stay probabilities overall increase following debriefing, and this also shows up in the perseveration parameter of the RL model, which increases following debriefing. We think this occurs because during the debriefing subjects are told that the reward probabilities at the two sides reverse only occasionally, and therefore, once they have detected a switch they expect the reward probabilities to be stable for an extended period. To examine this explicitly, we analysed whether the increase in perseveration as a whole due to debriefing correlated with a reduction in perseveration over the course of each block in the post debriefing data (Figure S7, right panels). The rationale is that if the increase in perseveration due to debriefing is due to expecting a period of stable reward probabilities after a block transition, then in subjects who show this effect strongly, we should also expect to see that, after debriefing, they are actually less perseverative late in blocks compared to early (whilst being more perseverative overall). This was indeed the case in healthy volunteers, who showed the debriefing effect on perseveration strongly, but not in either clinical group, both of which showed a significantly weaker (OCD group) or non-existent (mood and anxiety group) effect of debriefing on perseveration. This analysis is discussed in the manuscript as:

Results, lines 298-306*:* *Debriefing also increased how often subjects repeated choices independent of subsequent trial events, as reflected by a significant increase in the 'perseveration' parameter of the RL model (P<0.001; session by group interaction P=0.001; 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=-0.35, P=0.02; Figure S7).*

-- Authors focus on participants who did not use MB learning in session 3 (e.g. in Figure 4), but what did happen to participants who adopted MB after they received instruction?

It is an interesting question to understand what happens to behaviour on a task subjects have already learned a good model of from experience, when they are subsequently told the structure. Unfortunately, our data does not really allow us to speak clearly to this, because in the debriefing group on the fixed task, only 3 subjects had adopted model-based at session 3 as assessed by the likelihood ratio test. In these 3 subjects we do not see any obvious changes in behaviour as a result of debriefing, but we feel the sample size is so small that it is not worth reporting this in the manuscript. This analysis is shown below for the reviewers' interest:

Debriefing effects in the 3 subjects on the fixed task who were model-based at session 3.

Minor

-- Figure 4 caption: "in control"

-- Figure 5 caption: "in OCD and other mood disorders?"

We have corrected figure captions to clarify the target population for the data in each figure. We opted not to include ‘and other mood disorders’ given that this specific population was mostly intended as a control for the OCD population, both regarding the presence of anxiety/depression symptoms and the effects of medication.

Reviewer #3:

Remarks to the Author:

In the present manuscript, Castro-Rodrigues and colleagues used a modified version of the ‘two-step task’ task to assess contributions of model-based (MB) versus model-free (MF) systems to reinforcement learning in humans. Specifically, they tested to what degree humans deploy MB vs MF learning, first in the absence of any explicit instruction about the task structure, and then after fully explicit debriefing. Both healthy humans and individuals suffering from OCD (and another clinical control sample) were tested. The key finding according to the authors is that uninstructed behaviour was model-free, with model-based control only emerging over time in a subset of participants. The latter was less pronounced in the OCD group. All groups showed stronger model-based behaviour after receiving explicit debriefing on underlying task structure.

The manuscript is well written and addresses an important and interesting question, particularly given influential earlier reports that increased MF control may be a common feature of compulsive behavioural disorders (Voon et al., 2014). However, I am concerned whether the conclusions drawn by the authors are justified by the data and the study design.

We thank the reviewer for the recognition of the interest of the questions addressed. We have added several additional analyses to address the reviewers' concerns, as detailed in response to individual comments below.

1) One of the main findings is that uninstructed behaviour was dominantly model-free. This is a strong statement given the results and the analyses presented. In my view, what can be said with confidence is that subjects did not use the *true* model of the task - and this is the only test of 'model-based behaviour' the authors are presenting. Subjects could have used a completely different model of the task which would not be evident from testing to what degree they used the optimal model. Or indeed volunteers could have tried out different models of the task throughout the three sessions. For instance, as you present in Fig. S1, quite a proportion of people keep pressing invalid keys for a considerable period throughout at least the first session. This implies that it took subjects fairly long to understand even the most fundamental task characteristics. The authors write about there being no evidence of participants using task structure early on (line 337/338) - but given the above, everything else would be highly surprising!

Indeed, da Silva and Hare (which is also cited here) have shown that providing subjects with inaccurate models of the task evokes MB behaviour that can appear completely model-free. They also describe how behaviour which appears to be a hybrid of model-based and model-free could also be explained by a set of different algorithms (see also Collins & Cockburn, 2020). In this context, it would be important to see model comparisons that include competing models as well as model simulations on the eventual winning model.

This is an important issue, and we thank the reviewer for bringing it up and allowing us to address it in more detail. Our understanding of Silva and Hare's data is that they show that agents which are completely model-based, but have an incorrect model, can generate behaviour that looks similar to an agent that uses a mixture of model-based and model-free. As far as we can tell, all of their simulations from model-based agents with incorrect models still show a substantial transition-outcome interaction (the classical signature of model-based) but also show a significant main effect of outcome (normally considered a sign of model-free learning). This pattern is very different from that of our subjects at session 1, where we see a large main effect of outcome, but no influence at all of the transition outcome interaction. This is precisely the pattern expected for an essentially pure model-free strategy in which outcomes directly reinforce the preceding choice, and we are not aware of any incorrect-model-based strategies which generate behaviour that looks like this.

Nonetheless, we agree with the reviewer that, given the Silva and Hare data, it is important to explicitly test whether the types of incorrect-model-based strategies they consider might in fact explain our subjects' data. We therefore did an additional model-comparison on data from session 1 and session 3, in which we included the two incorrect-model agents proposed by Silva and Hare (termed 'unlucky symbol' and 'transition-dependent learning rate'), as well as the incorrect-model proposed below by the reviewer. Model-comparison indicated that all 3 of these incorrect-model agents fit data from both session 1 and session 3 less well than any of the traditional models we had previously considered. According to the reviewer's suggestion, we simulated behavioural performance of an agent using the best fitting model (mixture model) and observed it produced stay probability plots which were qualitatively similar to the collected data (Figure S6).

These results are presented in figure S2B and figure S6 (reproduced below), and discussed in the revised manuscript as:

Results, lines 169-174: *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 structure32, 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 (Figure S2b). We also simulated behaviour from the best fitting RL model and verified that it produced stay probability plots qualitatively similar to the experimental data (Figure S6).*

Supplementary figure S2b) Model comparisons including additional model-based agents with incorrect models of the task structure; one which believed state transitions were deterministic but volatile (IM-DV), one with transition dependent learning rates at the second-step (IM-TDLR) and one which believed that one first step option was unlucky and reduced reward probability at the second-step (IM-US).

Supplementary figure S6. Stay probabilities for best fitting RL model 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). Top panels show experimental data, bottom panels show behaviour simulated from the best fitting RL model, which used a mixture of model-based and model-free. Error bars indicated the cross subject standard error of the mean (SEM). In each group data was analysed separately for session 1 (blue graph), session 3 (red graph) and session 4 (gold graph).

2) An alternative model employed by participants could be that, rather than assuming a probabilistic, but fixed mapping of first-step choices to second-step states, participants

might assume this mapping to be fully deterministic, but highly volatile. Under such a model, following a rare transition, participants would infer that the underlying latent state has changed and assume that their first-step choice would now lead them to the state 2 they just observed. An agent using such a model would appear exactly like a completely MF agent. This could also be a potential explanation for the effects seen in the OCD group in figure 3 (A and D). It shows that, in OCD, there is an increase in MF control (Gmf) from session 1 to session 3 - and likewise an increase in the effect of outcomes. Again, I am not convinced that such a pattern is fully indicative of MF behaviour. As discussed above, the increase in P(stay) following rewards (irrespective of transition type) could as well be obtained from an inaccurate world model. Thus, such a pattern may actually indicate *increased* MB control in OCD!

We thank the reviewer for this suggestion. If we understand it correctly, the proposed agent is actually a special case of our model-based agent, because our model-based agent learns the transitions from experience, and if its learning rate for transitions (parameter) is set to 1, the agent believes that the most recently observed transition for a given action occurs with probability 1 when that action is selected. We would not expect this agent to generate a pattern of stay probabilities similar to that observed in our session 1 subjects, because model-based agents with a high learning rate for transitions generate strong loading on the transition predictor, in addition to the outcome predictor (Akam et al. PLOS CB 2015, Figure S5). Nonetheless we included it in the new model-comparison discussed above (figure S2B) but found that, as with the other incorrect-model agents, it fit the data worse than any of the traditional models that were considered.

3) I appreciate the motivation to simplify the original two-step task. However, by removing the choice alternatives at the second-step state, the task, at least to subjects, may lose its key characteristics as a multistep decision problem, but instead be perceived as a one-stage decision problem with probabilistic rewards (see my comment above). This might have influenced the perceived importance of forming a model and the consideration of the sequential structure of the task. Additionally, as the authors also note, the new design greatly reduces working memory load. Relationships between working memory capacity and task complexity on one hand and decision strategy on the other have been reported previously (Otto et al., 2013; Kim et al., 2019), which may also contribute to the apparent absence of MB behaviour.

Our motivation for removing the choice at the second-step was, as the reviewer points out, to simplify the task as much as possible, as we felt that the more complex the task was, the harder it would be for subjects to learn and use a model. Nevertheless, we appreciate the concerns raised by the reviewer, and have added to the discussion regarding the possibility that behaviour could be different in tasks with larger state-states:

Discussion, lines 528-532: *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.*

4) Furthermore, it was not clear to me why the authors opted for a fixed spatial-motor mapping instead of presenting two distinct visual stimuli with random mapping to top/bottom position on the screen. As the authors acknowledge in the discussion, such a fixed mapping of spatial position to effector may have further encouraged the use of a habitual/model-free/S-R strategy.

Again, the motivation was to simplify the task as much as possible, as we expected this would make it easier for subjects to learn a model of the task. As the reviewer notes we do discuss the possibility in the discussion that using fixed spatial-motor contingencies may affect the strategy used. We think it is equally possible that using discriminative stimuli whose position is randomised could make it harder, rather than easier, to use model-based RL, because action-outcome predictions and stimulus-outcome predictions, thought to be mediated by at least partially separate brain systems (Rudebeck et al. J.Neurosci 28.51 (2008): 13775-13785), would no-longer be aligned.

5) Modelling:

a) it would be re-assuring to see whether the fitted parameters can be recovered from simulated data generated using the fitted parameters (parameter recovery).

We now include a new parameter recovery analysis in which we test how well individual parameters can be recovered when we set the other parameters to their fitted values from either session 1, 3 or 4, and use the same number of simulated subjects as we have real subjects (Supplementary figure S8, reproduced below). This is discussed in the revised manuscript as:

Results, lines 307-311: *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 (Figure S8). 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.*

Supplementary figure S8. RL model parameter recovery. Test of the accuracy with which RL model parameters could be recovered from simulated data. Panels show mean and standard deviation of recovered parameters across 10 repeated simulation runs when the parameter under investigation was fixed at the specified 'true parameter value' and the other parameters were drawn randomly for each subject from the population level distributions fit to the specified dataset (top row- fixed task session 1, middle row – fixed task session 3, bottom row – fixed task debriefing group session 4).

b) Unlike in Daw 2011, there are separate weights (Gmb, Gmf) for the contributions of the MB and MF systems, respectively (rather than Gmf = 1 – Gmb). This seems plausible to me, as there may well be participants in which both MB and MF is low (e.g. purely stochastic or perseverative choice), but since this is also a - minor - departure from the original analyses, it would be helpful to briefly justify this in the text.

We thank the reviewer for this suggestion. We included a justification for the separate weights in the methods section of the revised manuscript (lines 732-734).

Methods, lines 704-706: *We used separate weights (G_mf, G_mb) for the influence of the model-based and model-free systems[30], 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].*

c) In figure 2F, the modelling results do not support the logistic regression - there is no difference in Gmf and Gmb between session 1 and 3.

The reviewer is correct that there is a discrepancy between the increase in the influence of the transition-outcome interaction on subsequent choice in the logistic regression between session 1 and 3 and the absence of a significant change in the model-based weight parameter of the RL model. This is discussed in the manuscript (lines 178-183) where we state:

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 subjects learned to use model-based RL; as model comparison for individual subjects between the mixture RL model and a simpler model-free RL model, indicated that only 15% of subjects (10/67) used model-based RL at session 3 (likelihood ratio test, threshold P=0.05).

We note that the maximum-a-posteriori fits for individual subjects (dots in figure 2F) indicate that a small number of subjects have approximately double the influence of model-based RL at session 3 compared to session 1, but this does not drive a significant change at the level of the population as a great majority of subjects show minimal change.

We also note that in addition to the model-based weight parameter, the value learning rate will also affect the transition-outcome interaction as it determines how much trial outcomes update the value of second-step states, which are used by the model-based system to compute the value of first step actions. This parameter did show a significant increase between session 1 and 3.

d) Bias parameter in the model: is my understanding correct that this indicates an action bias for the top circle? If so, why is it B = 1 for the high and B = 0 for the low action? Would this not mean that the model can have a bias for the upper circle, or no bias at all? Shouldn't B = –1 for the low option to also allow for a bias toward the lower option?

Apologies, this was not explained clearly enough in the methods. We have modified the relevant section to clarity how the bias and perseveration parameters work, it now reads (lines 685-694):

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

where and are parameters controlling, respectively, the strength of influence of model-free and model-based action values on choice. is a parameter controlling the strength and direction of choice bias, is a variable which takes a value of 1 for the up action and 0 for the down action. Since it is the difference in values between up and down actions that determines choice, positive values of therefore generate a bias towards the up action and negative values towards the down action. is a parameter controlling the strength and direction of choice perseveration, is a variable which takes a value of 1 if action was chosen on the previous trial and 0 if it was not. Positive values of therefore promote repeating the previous choice while negative values promote switching.

e) Decreases in the eligibility trace parameter lambda correlate, across subjects, with increase in MB control - is this naturally arising because as lambda --> 0, there is no more update in the MF system? In other words, is it possible, for a given subject, that there are two (probably very similar) local optima in the parameter space, at low (high) values for lambda and high (low) values for Gmb?

The eligibility trace parameter does not affect the relative influence of model-based and model-free values on choices, it just changes the way that model-free values for first-step actions are updated. Specifically, when is 0 the update to first-step action values depends only the value of the second-step reached, and when is 1 the update depends only on the trial outcome, with intermediate values of determining the mixture of these two influences on the update. We therefore would not expect problems in accurately fitting as long as there is a model-free influence on choices (as only affects updates to the model-free first step action values). We verified this empirically with the new parameter recovery analysis (figure S8, discussed above), confirming that we are able to accurately recover the values of both the model-free weight and when the other parameters of the simulated data match those fitted to the real data from either session 1,3 or 4.

6) In figure 2F, the modelling results indicate no difference between MB and MF contributions to behaviour. This is at odds with the logistic regression results presented in 2D (and the p(stay) in 2C) which indicates a clear transition X outcome interaction. This again raises the issue to what extent the model with the fitted parameters is able to recapitulate the actual pattern in subjects' choice behaviour (related to my question above regarding parameter recovery)

If we understood correctly, we think the reviewer might be reiterating the point here that they made in comments 5a and 5c. We now included a parameter recovery analysis and further discussed this issues in lines 307-311 and lines 178-183, as well as in our above responses to comments 5a and 5c.

7) In the analyses investigating the effects of providing the full task structure, it seems that out of the 57 healthy individuals not showing MB behaviour yet, n = 41 were assigned to the debriefing group whereas only n = 16 were not debriefed. Why were they evenly allocated to both groups? Comparing the proportion of subjects acquiring MB control appears odd given the highly unbalanced group size?

Initial experiments were conducted among healthy volunteers recruited in Lisbon, with two variants of the task (Fixed and Changing task). 82 participants were randomized between the two versions of the task and, within each task, some participants were debriefed between sessions 3 and 4 (17 performing the Fixed version and 16 performing the Changing version of the task), while the remaining subjects were not debriefed (23 in the Fixed version and 26 in the changing version). From the 23 non-debriefed subjects in the Fixed version, 16 were not showing model-based at session 3.. Given the results from this first sample, in healthy volunteers subsequently recruited in New York (n=27), as well as all clinical samples in both sites, only the Fixed task was used, and all participants were debriefed. Thus, the n=41 subjects who were not model-based at session 3 were 25 of the 27 healthy volunteers recruited in NY and 16 of the 17 healthy volunteers who performed the Fixed task with debriefing in Lisbon. As a result, the sample is unbalanced towards the participants performing the Fixed task and that were debriefed. To address potential problems resulting from the unbalanced nature of the debriefed and non-debriefed samples, as mentioned in the results section, analyses were repeated including only participants recruited in Lisbon, which did not affect the results. Furthermore, we tested differences in learning or debriefing effects between the Lisbon and New York debriefing groups, and did not find any significant differences (Table S3).

This is now clarified in the methods section, lines 633-635, where it now reads: *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.*

8) Why are people diagnosed with a wide variety of mood and anxiety disorders put together as one group? These disorders are characterized by several different symptoms which could also be expected to influence MB/MF control of behaviour in different ways. I guess it also does not serve to "investigate potential contributions of medication or unspecified mood and anxiety symptoms" for the same reasons. Further details on psychotropic medication would be needed to evaluate their comparability with respect to medication.

We decided to include a group of participants with different mood and anxiety disorders because there is a very high comorbidity between OCD and mood disorders (such as depression) and anxiety disorders (such as generalized anxiety disorder). In fact, for each diagnosis present in any group (depression, bipolar disorder, dysthymia, generalized anxiety disorder, panic disorder, agoraphobia, social anxiety disorder, post-traumatic stress disorder, anorexia nervosa, trichotillomania and hoarding disorder), we tested if its proportion was different between the OCD group and the mood and anxiety group, and we did not find significant differences for any diagnosis (P's>0.3; chi-squared or Fisher exact test), except for obsessive-compulsive disorder, which was an exclusion criteria for the mood and anxiety disorders group (data not shown). Also, to the best of our knowledge, there is no consistent evidence for imbalance in model-based/model-free control in any specific mood or anxiety disorder. Nevertheless, the choice of a group of several disorders as a 'control', rather than a specific disorder, reduces the risk of observing effects dependent on the 'control' condition, rather than effects dependent on the disorder of interest (OCD). Importantly, scores in assessments of anxiety and depression symptom severity were equivalent between the two clinical groups, that were only different according to severity of OCD symptoms. This suggests that this group can in fact allow for control of the effects of unspecific mood and anxiety disorders, while studying the more specific effects of OCD symptoms.

Regarding medication, we classified it in classes (SSRI, tricyclic antidepressant, second-generation antipsychotic, first-generation antipsychotic, mood stabilizers, benzodiazepines, other antidepressants and other medications) and we did not find differences between groups in the use of any class of medication (P's>0.2; Chi-squared or Fisher's exact test) (see Table below). It is also important to mention that while the clinical groups recruited in Lisbon were medicated, the clinical groups recruited in New York were unmedicated. As reported in the original manuscript, we did not find differences between Lisbon and New York groups in any behavioural or psychometric measure.

Table r1 – Psychotropic medication in each group. Data was compared between groups using Chi-square test or Fisher's exact test (when samples were very small). OCD = Obsessive-compulsive disorder; MA = Mood and anxiety disorders group; SSRI = Selective serotonine reuptake inhibitor; TCA = Tricyclic antidepressant; SGA = Second-generation antipsychotic; FGA = First-generation antipsychotic; BZD = Benzodiazepine).

We now present this in the manuscript, lines 110-114, where it now reads: *For each diagnosis present in any group, we tested if its proportion was different between the OCD group and the mood and anxiety group, and we did not find significant differences for any diagnosis (P's>0.3; chi-squared or Fisher exact test), except for OCD. Regarding medication, we classified it in classes and we did not find differences between groups in the use of any class of medication (P's>0.2; Chi-squared or Fisher's exact test).*

9) RT at 2nd stage are faster in common compared to rare transitions, and this effect is present already in session 1 (but seems to become larger in session 3). Is this an indication that subjects have learnt the state transition probabilities in session 1 already?

The reviewer is correct that reaction times at the second-step are faster following common than rare transitions even in the first session where choices are model-free. We think this likely reflects motor systems learning to predict, and hence prepare, upcoming actions before higher-level decision-making systems have learnt to use state predictions to guide choices. This is discussed in the manuscript as:

Results lines 143-147: *Although state transitions in session 1 did not influence subsequent first-step choices, 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; t66=7.81, P<0.0001, 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.*

Results lines 184-192: *Key-press reaction times at the second-step became faster overall between session 1 and 3 (main effect of session P<0.0001, repeated measures ANOVA), but this was more pronounced following common than rare transitions (session-transition interaction P=0.008; Figure 2e). Additionally, by session 3 the rate of invalid key presses was significantly higher following rare than common transitions (P=0.002, Wilcoxon signed rank test, Figure S1). Therefore, both choice-based and implicit measures showed evidence of learning about the transition structure between session 1 and 3. Model-based choice was significantly correlated across subjects with the rare-common reaction time difference at both session 1 & 3 (P<0.001, figure S3), suggesting interaction between learning at motor and cognitive-levels, thought this appeared to be driven by the minority of subjects whose choices were more model-based.*

See also our response to reviewer 1's question about this.

10) Before debriefing participants, did you question them about what kind of strategy they used? While this would probably hard to quantify, it might at least give some qualitative hints as to what beliefs they had acquired (and what models of state space they applied) during the 900 uninstructed trials.

We thank the reviewer for this question. Prior to informing subjects about the task structure we did indeed assess their beliefs, using a pen-and-paper questionnaire given at the end of session 3. We had not included this data in the previous version of the manuscript because we were unable to access the documents with the NY groups responses due to COVID-19 restrictions preventing access to the building. We have now been able to analyse this data for all subjects and include the results in new Supplementary information (lines 46-99).

Participants were required to answer four questions. For the first two questions, they were asked to provide open-ended answers. Afterwards, the same questions were asked in a multiple choice format.

The questions were asked after showing the participants the following image in a sheet of paper:

Specifically, the first question was: "If you press the upper arrow key when this screen is being shown, what is most likely to happen next?". The second question was: "If you press the lower arrow key when this screen is being shown, what is most likely to happen next?"

After the participants wrote their answers, the researcher collected the first sheet of paper and provided them a second sheet of paper, which also contained the same image at the top and the same two questions. However, instead of writing an open answer, participants were asked to choose one of four different options for each question:

a) The left side circle will turn yellow

b) The right side circle will turn yellow

c) A coin will appear on the left side

d) A coin will appear on the right side

To classify the open answers that participants gave as correct or incorrect, we created a set of criteria that the answers had to fulfil in order to be considered correct. According to these criteria, each answer had to contain at least one of the following elements: a) "the right circle would turn yellow / light up / be highlighted"; b) "the yellow circle moves [from the top circle] to the right circle"); c) "most of the times, the right circle would turn yellow, although in a minority of times the left circle turns yellow". Two independent raters (PCR & AM) assessed open answers independently. Then, when in disagreement, they discussed and reached a consensus if those answers should be considered correct or incorrect. The rate of concordance between independent raters was 92%.

Considering all participants together, we found no association between correct or incorrect answers and pre-debriefing behavioural measures, either in open answers (-1.7<t's<0.2; P's>0.09; independent sample t-test's) or in multiple-choice questions (-1.2<t's<0.1; P's>0.2). However, we found that subjects who gave correct multiple-choice answers had a higher influence of model-based action values on choice after the debriefing (t = -2.66; P=0.009; independent sample t-test).

Concerning open answers that did not fill the criteria for being considered correct, we identified a frequent type of wrong belief/model, specifically that the second-step circle is random (~25% of open answers). Regarding the remaining incorrect open answers, they consisted of different types of answers each occurring with a small frequency (<10%), such as ignorance of the second step ("immediately after the first-step choice, a coin may appear or not") or an incorrect transition model (common transitions identified as rare & rare transitions identified as common). As the first type of wrong model (random second-step) was particularly frequent, we divided subjects in three groups (correct model, random second-step, other wrong models) and tested if their pre-debriefing behaviour was different. We did not find significant differences between groups in terms of behavioural strategy, either in logistic regression or model fitting analyses (F’s<2.0; P’s>0.14; one-way ANOVA).

When comparing different clinical groups, we found that while individuals with OCD had a smaller proportion of correct open answers (12/39) than healthy volunteers (23/43), this difference was not statistically significant (χ2 = 2.1, P=0.1, chi-squared test). We also did not observe differences between OCD and healthy volunteers in their proportion of correct multiple-choice answers (χ2 =0.45, P=0.5). Individuals with mood and anxiety disorders had the same proportion of correct/incorrect answers that healthy volunteers, both in open answers (χ2 =0.2, P=0.7) and in multiple-choice (χ2 =0.03, P=0.9).

Our findings complement previous data provided by other groups using an operant conditioning outcome devaluation paradigm in which OCD subjects did not differ from healthy volunteers in their capacity to provide explicit assessment of the contingencies governing action and outcome, although their behaviour did not reflect these (Gillan et al., 2014; Gillan et al., 2015). However, in those studies outcomes were aversive and the experimental induction of habits in humans using operant conditioning paradigms has been questioned (De Wit et al., 2018; Robbins et al., 2019).

11) It is unclear to me why the neutral blocks existed? What was their purpose?

The original two-step task (Daw et al. Neuron, 2011) used reward probabilities which drift as random walks on the range 0.25 - 0.75, so most of the time they are pretty close to neutral. This has the advantage of reducing correlations across trials, which can otherwise complicate analysing how events on one trial affect subsequent choices (see Akam et al. PLOS Comp. Biol. 2015 for a detailed discussion). However, it also means there is very little contrast between good and bad options, and means that model-based RL does not yield a significantly higher reward rate than model-free (Akam et al. PLOS Comp. Biol. 2015, Kool et al. PLOS Comp. Biol 2016). We felt it was important that there was a significant contrast between good and bad options to promote task engagement and ensure that model-based RL yielded higher reward rates, hence having 80-20 non-neutral blocks which only transitioned once the subject started choosing the correct option, but incorporated neutral blocks as well to try and reduce cross trial correlations, at least to some extent.

Minor:

1) In the introduction (Line 79 onwards) you write: "However, no studies have explored behaviour on multi-step tasks in the absence of information about task structure, in

either healthy or clinical populations" <-- I can think of one study (Gläscher et al. 2010, Neuron) that used Tolman-style latent learning in an abstract multi-step T-maze. I think this is also without explicit instruction and likewise looks at the contribution of MF and MB systems over time.

We thank the reviewer for pointing out this relevant prior work. It is true that subjects in Gläscher et al. learned the task structure from experience (during the latent learning phase) then subsequently used this knowledge in the rewarded phase to guide choices. We have modified the introduction to state that 'few' rather than 'no' studies have explored this question, and now discuss the Gläscher et al study explicitly in the discussion as:

Lines 534-539: *Another question is how information given to subjects 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, subject 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*[14]*.*

2) In the discussion (#341): "When learning from experience, individuals with OCD were impaired in their use of model-based control and biased towards a model-free strategy." This makes it sound as if this was only true for OCD, but it likewise applies for healthy controls.

We have clarified this in the revised manuscript, lines 379-381, where it now reads: *"When learning from experience, individuals with OCD were impaired in their use of model-based control and were more biased towards a model-free strategy, as compared with healthy volunteers."*

3) Line 283/285: "Increased use of model-based RL after debriefing was confirmed by model fitting (Figure 5d), which showed increased influence of model-based action values on choice" <-- I guess this should be Figure 5e?

Thanks, fixed in revised manuscript.

Decision Letter, first revision:

10th June 2021

Dear Professor Oliveira-Maia,

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

Thank you for submitting your revised manuscript and for all your work on the revision. I appreciate in particular that you collected additional data to respond to the referees' concerns.

Although your manuscript has been revised in response to reviewer comments, it does not fully comply with our editorial policies and formatting requirements. In particular, you must revise the presentation and interpretation of statistical tests before we can proceed to peer-review.

In terms of presentation, this includes provision of full statistics (please see the attached document) for all parametric tests; provision of equivalent information for non-parametric tests is likewise necessary (especially a measure of effect size and a measure of the effect size's confidence/credibility interval where possible).

In terms of interpretation, please not that we do not accept absence of evidence for an effect as evidence of absence; moreover, we do have very clear guidelines as to how null-effects (effects where p > alpha) may be described. To provide a few examples:

'During session 1 there was no difference in the average rate of invalid key presses at the second-step following common vs rare transitions (P=0.19, Wilcoxon signed rank test, Figure S1).'
should read:
'During session 1 there was no statistically significant evidence for a difference in the average rate of invalid key presses at the second-step following common vs rare transitions (P=0.19, Wilcoxon signed rank test, Figure S1).' [You must also provide full statistics, not just the p-value for non-significant results]

'During session 1, stay probability was strongly influenced by trial outcome (P<0.001, permutation test), but not by state transition (P=0.3) nor the transition-outcome interaction (P=0.1; Figure 2 a, d).'
should read:
'During session 1, stay probability was strongly influenced by trial outcome (P<0.001, permutation test). There was no statistically significant evidence for an influence of state transition (P=0.3) or the transition-outcome interaction (P=0.1; Figure 2 a, d) on stay probability'. [You must also provide full statistics, not just the p-value for significant and non-significant results]

'As in the self-paced task, initial behaviour was consistent with model-free control, with a main effect of trial outcome on stay probability (P < 0.001, permutation test) in session 1, but no effect of transition (P=0.23) or transition-outcome interaction (P=0.92) (Figure S4a,b)'
should read:
'As in the self-paced task, initial behaviour was consistent with model-free control, with a main effect of trial outcome on stay probability (P < 0.001, permutation test) in session 1. There was no statistically significant evidence for an effect of transition on stay probability (P=0.23) or statistically

significant evidence for an interaction between transition and outcome(Figure S4a,b)'. [You must also provide full statistics, not just the p-value for significant and non-significant results]

'However, the session by transition interaction did not reach significance (P=0.2), indicating that for this group the effect of experience on reaction time did not differentiate between common and rare transitions.'
should read:
'However, the session by transition interaction did not reach statistical significance (P=0.2)'. [No interpretation of null-effects derived from NHST is permitted, hence, they cannot indicate anything in our pages].

'[...]unlike in healthy volunteers, did not correlate with decreases in perseveration from early to late in blocks after debriefing (r=-0.09, P=0.5; Figure S7)'
should read:
'[...] in patients we found no statistically significant correlation with decreases in perseveration from early to late in blocks after debriefing (r=-0.09, P=0.5; Figure S7)'. [Please avoid logically contrasting significant effects and the absence of significant effects. Please note also that you must report full statistics for correlation analyses, too, including for ns findings.]

These are some examples, but the list is much more comprehensive.

In brief, no interpretation or analysis choice (i.e. treating groups as the same) may be based on a non-significant result in null-hypothesis significance testing.

Before we can send the manuscript back to our reviewers, we ask that you revise it to ensure that it complies fully with our policies and is formatted according to our requirements. I have attached another copy of our checklist that exemplifies our formatting and policy requirements. If you are uncertain as to how to address any of the points in the checklist, please don't hesitate to contact me.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you in advance for attending to these requests and I look forward to receiving your revised manuscript.

Sincerely,

Marike Schiffer

Marike Schiffer, PhD
Acting Chief Editor
Nature Human Behaviour

Author Rebuttal, first revision:

Reviewer #1

Remarks to the Author:

The study by Castro-Rodrigues, Akam et al. looks at the contribution of model-based and model-free control to action selection in a simplified version of the classic 'two-step' task. In particular, they examine whether subjects' behaviour is initially model-based or model-free when subjects receive minimal instruction on the structure of the task. This is potentially of interest because most previous studies have given explicit instruction about the structure of the task, and a recent report from Todd Hare's lab has found that when these instructions are sufficiently clear then humans tend to be exclusively model-based.

They find that most subjects show behaviour that is initially model-free, not model-based. This behaviour persists over several days of repeated exposure to the task, even though subjects show implicit measures of learning the task structure. When subjects are explicitly instructed about the structure of the task, their behaviour becomes predominantly model-based. Their two central conclusions from this are: (i) explicit task structural knowledge determines human use of model-based RL, and (ii) this is most readily acquired from instruction rather than experience.

In addition to this, the authors explore variation of these behaviours between patients with OCD and other disorders. While there are some differences in the degree to which OCD patients use model-based control, in that uninstructed behaviour is even more biased towards model-free behaviour in these patients, the overall pattern of behaviour is similar to healthy controls (in other words, the patients become more model-based with instruction).

The paper overall provides an interesting contribution to the study of this task, and in some ways complements the recent study by Silva and Hare, by showing what behaviour is like in the "naïve" state of not knowing anything about the task structure. Overall, I thought that the data were well analysed/modelled and clearly presented.

We thank the reviewer for the positive assessment of our manuscript.

On the other hand, I was left slightly concerned about whether the two central conclusions mentioned above can really be generalised to all learning tasks where the structure of the task is unknown, or whether the conclusions are actually far more narrow and only apply to the situation studied here. I outline these concerns below.

1. My first concern is that subjects performed this task exceedingly quickly, without much deliberation at all (reaction times were typically below 400ms, and subjects completed 1200 trials per session). This leads to a concern that participants were optimising speed over accuracy in completing the task – in which case, a model-free controller might well be the optimal strategy. The task appears to be entirely self-paced (although I couldn't find details on the duration of reward presentation, duration of inter-trial interval, duration of interval between light 1 being extinguished and light 2 appearing, etc. in the methods), and subjects may well not be trying to optimise number of rewards but instead trying to complete the task reasonably quickly. One possible way to address this would actually be to collect a small additional dataset where the number of trials is brought down considerably (e.g. 300 trials per session), and the pace of the task is slowed right down (e.g. each light presentation event lasts 1 second before the subject can respond, and there is a 1 second ITI). In such circumstances, would subjects still appear entirely model-free until they were instructed otherwise? My suspicion is that they would not. Although I accept that the timings of the task are unchanged in session 4 (the instructed session), the instructions are essentially telling participants that they need to pay attention to these features when performing the task (and indeed, they become more deliberative on rare transitions in these circumstances (figure 4d).

We agree with the reviewer that, in principle, due to the self-paced nature of the original task, and overall fast reaction times exhibited by subjects, they might have been optimising speed over accuracy and hence opting for a model-free strategy. We directly addressed this in the revised manuscript by following the reviewer's suggestion and gathering an additional dataset in which we slowed down the pace of the task by introducing a 1 second delay between options being highlighted and being active for selection (cued by a colour change from pale to bright yellow), at both the first and second step, in addition to a 1 second inter-trial interval. In this dataset, 20 newly recruited healthy volunteers completed a total of 600 trials each, split into 3 sessions of 150 trials prior to debriefing, and an additional session of 150 trials after debriefing (the original task used 4 sessions each of 300 trials). Despite the slower pace and hence more time for deliberation, the new data strikingly recapitulated our original findings, including:

- A large influence of trial outcome but no statistically significant effect of transition-outcome interaction at the first session, consistent with a model-free strategy (Figure S4, reproduced below).

- A small increase in transition-outcome interaction by session 3, consistent with increased use of model-based RL with experience (Figure S4b). As in the original task, this appears to be driven by a minority of subjects, as a likelihood ratio test on session 3 data supported a mixed model-based + model-free strategy over a model-free only strategy in only 3 among the 20 subjects.

- A large increase in the use of model-based RL after receiving explicit information about task structure. As in the original task, this was accompanied by a decrease in the RL model eligibility trace parameter, which we think reflects changes to subjects' representation of the task-state space affecting model-free value updates (Figure S5, reproduced below).

Overall these data confirm our main results that, in unfamiliar domains, subjects initially rely on a model-free strategy, that they are surprisingly slow to learn to use a task model from experience, but learn much more readily from explicit description.

These data are detailed in the revised manuscript as:

Results 206-224: *A possible reason why model-free control might predominate is that subjects 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). Subjects 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 (Figure S4a,b). Also, similarly to the self-paced task, the effect of transition-outcome interaction on stay probability increased between session 1 and 3 (95% CI of null hypothesis test statistic [-0.35,0.34], coefficient change=0.45, P=0.005, permutation test), as assessed by the logistic* *regression (Figure S4b), 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 (Figure S4d), 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 subjects.*

Results 294-296: *Similar effects of debriefing on the transition-outcome interaction were found among the 17 healthy volunteers performing the slow-paced version of the task that were not using model-based RL significantly in session 3 (Figure S5a-c).*

Supplementary figure S4. Learning effects in slow-paced task. a) Stay probability analysis. **b)** Logistic regression analysis of stay probabilities. **c)** Reaction times after common and rare transitions in session 1 and 3. **d)** Comparison of mixture model fits between session 1 and session 3. RL model parameters: MF, Model-free strength; MB, Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.

Supplementary figure S5. Effects of explicit knowledge in the slow paced task. **a)** Per-subject likelihood ratio test for use of model-based strategy on session 3 (left panel) and session 4 (right panel). **b)** Stay probability analysis. **c)** Logistic regression analysis of stay probabilities. **d)** Reaction times after common and rare transitions in session 1 and 3. **e)** Comparison of mixture model fits between session 1 and session 3. RL model parameters: MF, Model-free strength; MB, Model-based strength; αQ, Value learning rate; λ, Eligibility trace; αT, Transition probability learning rate; bias, Choice bias; pers., Choice perseveration.

2. More broadly, it is unclear whether the general conclusion that participants need explicit task knowledge in order to use model-based RL is really true – or whether it is only true in a task with a very limited state space, like the present task. Model-based control seems far more important in tasks with much larger state spaces than the one considered here. So while the present study provides interesting insights into this particular task, and provides an interesting comparison with the many previous studies that have used this task, it is unclear whether it can be used to draw a broader conclusion about the use of model-based versus model-free control per se.

We agree that it is possible that the size of the task state space might affect use of model-based RL. It is worth noting that the task did incentivise the use of model-based RL, as shown by the positive correlation we observed between measures of model-based choice and total rewards obtained, demonstrating that use of model-based RL is important in our task. We also note that, while larger state spaces might increase payoffs for using model-based RL, they also make both model-learning and planning substantially more difficult, so it is not obvious that model-based

RL would necessarily be utilised more or earlier in larger state spaces. Nonetheless we agree it is important to be transparent about this limitation of our study, which we now discuss in the revised manuscript file as:

Discussion, lines 590-594: *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.*

We have also softened the language in the title and abstract to state that explicit knowledge is ‘a’ rather than ‘the’ primary determinant of human model-based action.

3. Figure S3 shows that in about 10% of trials, participants make invalid presses at the second stage. It would be interesting to know if these invalid presses reflect implicit learning of the transition probabilities (i.e. whether they are more common on rare transitions than common transitions), and if so, when this develops across the different sessions (and whether it is at all predictive of the degree of model-based control, or the strength of the common/rare reaction time effect in figure 2e). The fact that subjects appear to be learning the model structure implicitly makes me feel more strongly that if participants were forced to become slower in their responses (comment 1, above), they would likely become more model-based.

We thank the reviewer for raising this important question, addressed via a new analysis, added to figure S1, comparing the rate of invalid presses at the second-step (left or right actions), following common and rare transitions. In the first session the rate of invalid presses following common and rare transitions were not significantly different, but by the third session subjects made significantly more errors following rare transitions. This provides further evidence for implicit learning of task structure, in addition to the previously reported reaction time differences following common vs rare transitions, which are significant even in the first session.

We have also added an additional supplementary figure (Figure S3, reproduced below) looking at correlations, across subjects, between measures of model-based choice (model-based weight from the RL model and transition-outcome interaction from the regression) and measures of implicit learning of the transition structure (differences in second-step reaction times and rates of invalid responses following common vs rare transitions). Interestingly, choice-based and implicit measures of task structure learning, reaction-time in particular, were significantly correlated at all time-points evaluated (sessions 1,3 and 4). This appears to be driven, at least in sessions 1 and 3, when subjects are learning only from experience, by the handful of subjects whose choices are more model-based.

Taken together, we think these data support a partial dissociation between task structure learning at the motor and cognitive level, with motor-level learning occurring faster and more consistently across the population, as evidenced by the robust reaction time effects at session 1 when choices are model-free, but with some interaction between the levels evidenced by the cross-subject correlations.

We agree with the reviewer that evidence for implicit task structure learning further motivates the control experiment we performed in response to comment 1 above, in which we forced subjects to do the task at a slower pace. As discussed above, choice behaviour in this slow-paced task was similar to that in our original data. We think these data are consistent with a dissociation between task structure learning at motor and cognitive levels, in that robust implicit learning of task structure, as evidenced by motor performance, does not necessarily imply that cognitive systems have a usable task model.

We discuss these data in the revised manuscript as:

Results lines 148-153: *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; t66=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.*

Results lines 195-205: *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, Figure S1). 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 subjects with the rare-common reaction time difference at both session 1 (r=0.57; P<0.001; Pearson's correlation, figure S3) and session 3 (r=0.69; P<0.001), suggesting interaction between learning at motor and cognitive-levels, though this appeared to be driven by the minority of subjects whose choices were more model-based.*

Discussion, lines 543-550: *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[56]. Intriguingly, implicit and choice-based measures of task structure learning were correlated across subjects, even at session 1, suggesting that this dissociation is only partial, with cognitive task models potentially informed by earlier motor-level learning.*

Supplementary figure S3. Correlation between choice and implicit measures of task structure learning. Correlation across subjects between different measures of task structure learning in the fixed task at sessions 1, 3 and 4. Two choice-based measures were used; the model-based weight parameter from the RL model fit and the transition x outcome predictor loading from the logistic regression, and two implicit measures; the difference in second-step reaction times following common vs rare transitions, and the difference in the rate of invalid second-step responses (e.g. pressing left when the right state was active) following common vs rare transitions. Points show individual subjects. Lines show linear fit with 95% confidence interval on fit indicated by shaded region. In both session 1 and 3, when subjects are learning only from experience, both measures of model-based choice are correlated with the rare-common reaction time difference.

4. In a version with changing probabilities, which was run as a pilot study, the authors conclude that subjects were not really able to learn anything at all about the task structure. But a comparison of supplementary figure S6a and figure S6f suggests that a small subset of subjects could learn the task structure, and the proportion of participants didn't really differ between instructed vs. non-instructed versions.

The reviewer is correct that some individual subjects on the changing task did show evidence of model-based RL in the last session (session 4), as assessed by the likelihood ratio test. The reviewer is also correct that informing subjects about the structure of the changing task did not significantly increase the use of model-based RL, since the proportion of participants using model-based RL on session 4 was the same independent of debriefing (2/12 vs. 4/24). Also, the number of participants that started using model-based RL on session 4 (6/36) was similar to those using model-based RL on session 3 (6/42). The lack of effects of debriefing presumably reflects the fact that participants did not understand the debriefing, or felt the task was too complex to make the effort. This was not the case in the fixed task, where debriefing dramatically boosted use of model based RL. This was one of the reasons why we choose the fixed task for the majority of our experiments. The other reason was that, taking the population performing the changing task as a whole, we saw no evidence of a significant increase in the use of model-based RL during the uninstructed sessions, i.e. we did not see the increase in transition-outcome interaction that we saw in the fixed task. We have clarified this in the revised manuscript as:

Results, lines 119-122: *The Changing version proved too complex for most subjects, 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 Supplementary information for details), so we subsequently focused on the Fixed task, which is used for all data and figures in the main text.*

Supplementary information, lines 8-9*: The proportion of subjects for whom a likelihood ratio test indicated model-based RL was being used in session 3 (6/42) was similar to that observed in the Fixed task (10/67).*

Supplementary information, lines 24-28: *These data suggest that, while model-free RL was dominant for the non-instructed sessions of both tasks, the dynamically changing action-state transition probabilities in the Changing task further reduced the ability to learn a model-based strategy. This is consistent with uncertainty-based arbitration between model-based and model-free control[33,4], as the changing transition probabilities would be expected to increase uncertainty in the model-based system.*

Reviewer #2:

Remarks to the Author:

In this manuscript, Castro-Rodrigues and colleagues aim to study the effect of extensive task instruction (or the lack of it) on one of the most commonly used experimental paradigm related to model-based vs. model-free learning in healthy individuals as well as individuals with OCD and other mood and anxiety disorders.

The authors use a simpler version of the two-step learning task by Daw et al 2011, by presenting all choice options with four circles at once and removing the second-step choice while keeping the main components of the task in terms of state transitions.

The main finding is that in the absence of explicit instruction about the structure of the task, normal and clinical groups predominately adopt a model-free approach while a minority of subjects slowly moved toward model-based approach over time. Moreover, they only found a small difference between healthy and clinical groups in terms of the learning strategy.

It is refreshing to see that finally there is an investigation into the effect of explicit/extensive task instructions in the original version of a task that has caused a lot of controversy about model-based vs. model-free learning, and how much of previous

results were dependent on such task instructions. I think this is an important study which is executed and analysed well, and has important implications for studying learning in healthy and clinical populations. The manuscript is clearly written and the main claims of the study are supported by the results presented.

We thank the reviewer for the positive assessment of the manuscript.

I mainly have a few clarifying questions about the methods/results, some of the side claims and broader implications of the study, and finally some suggestions for additional analyses discussions.

-- I wonder why the authors did not use the original task of Daw et al. I could be wrong but I assume because it is almost impossible to perform the original task without explicit/extensive instructions. If so, what we have learned from many studies using the previous task? There is an interesting paper by Collins and Cockburn (Nat Rev Neuroscience 2020) that is worth discussing.

The reviewer is correct that we did not use the original task because we predicted that it would be too complicated for participants to learn uninstructed. We intended to simplify the task as much as possible while preserving the structure necessary to dissociate model-based and model-free RL, in order to optimize the chance of observing model-learning.

While we think our findings provide important insights into when humans do and do not use model-based RL, we do not think they invalidate findings with the original task. Rather, they help to understand where those finding are likely to apply and where they may not. Specifically, our data speaks to situations where subjects must learn from experience to act in an unfamiliar domain, and show that behaviour in these circumstances is very different to that in situations where subjects are given detailed information about task structure. Both of these situations occur in real-life decision making, and are hence important to understand, though we think the uninstructed situation is under-represented in laboratory studies. Our study contributes to clarifying processes associated with uninstructed learning and the effects of instruction.

We agree the Collins and Cockburn paper is worth discussing and do so as:

Discussion, lines 584-590: *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, 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 subjects 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.*

-- Authors mention that the Changing version of the task was too complex and that is why they focused on the Fixed version in most of their analyses (although they provide results based on the Changing version as well). Considering that blocks would end in Fixed version of the task, could not participants use this signal to learn faster? For example, did the time to reach criterion become shorter over block of trials?

Following the reviewer's suggestion, we quantified block length distributions on the fixed and changing version of the task during sessions 1-3, when subjects are learning from experience, and found that they were very similar (Fixed task 55.7 ± 32.1 trials, Changing task 55.7 ± 32.2 trials, mean ± SD). In these sessions most subjects are using model-free RL, and, in both tasks, model fits to these initial sessions indicate that the model-free update of first-step action values was primarily driven by the trial outcome (rewarded or not) rather than the value of the second-step state (this is indicated by the model's eligibility trace parameter λ being close to 1). For a model-free agent which updates the first step action values only from the trial outcome, irrespective of the second step state reached, the fixed and changing tasks are identical. Independently of whether the reward probabilities or transition probabilities reverse, the update required to the first-step action value is the same.

Importantly, our assessment that the Changing version was too complex was strongly influence by the effects of debriefing in both task versions. Specifically, debriefing significantly increased the proportion of individuals developing model-based RL in session 4 for the Fixed task (52.1% for the debriefed group and 6.3% for the non-debriefed group), but not for the Changing task (16.7% both for debriefed and non-debriefed groups). This has been clarified in the revised manuscript at lines 119-122, which read:

The Changing version proved too complex for most subjects, 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 supplementary information for details), so we subsequently focused on the Fixed task, which is used for all data and figures in the main text.

Regarding the reviewers' question about whether subjects got faster at adapting to reversals with task experience, we looked at this by examining the choice probability trajectories following reversals (see figure below) but did not observe significant changes between session 1 and 3 in either task (permutation test for difference in exponential fit P>0.17).

Also considering this task structure would not make more sense to analyse data across blocks instead of arbitrary sessions of 300 trials? Authors show one aspect of behaviour over time (in Figure S1) but not the main analyses (e.g. probability of stay, RT, etc.). I think including such analyses can be more informative, specially about transition from MF to MB.

We thank the reviewer for this suggestion, which we considered very carefully. On balance we think that analysing the data in blocks is probably not straightforward, because we do not see a way to look at the time-course of learning across blocks which does not either a) exclude a large number of subjects, or b) incur major biases due to including different subjects at different time points. This is because the number of blocks completed by subjects for each of the first 3 sessions varies over a wide range (shown below for the fixed task). If we divide the data based on block number, then either we will have many more subjects in the early data, or we will have to exclude a substantial quantity of either trials or subjects to balance the data. This would be further complicated by the fact that the number of completed blocks will be correlated with their behavioural strategy.

We think that grouping the data into early and late learning by session number, and hence trials completed, is a reasonable and principled approach to categorising the stage of learning. The choice of 300 trial lengths for sessions was, as the reviewer correctly points out, arbitrary, but given that this is what we used for the experiment it is a natural choice to also use for the analysis. It is also worth noting that we obtained qualitatively similar findings in the slow paced control experiment in which the length of sessions was 150 trials.

-- There are a few generally untested ideas about the utility of MB and MF RL, which appears in the introduction of this manuscript as well (lines 44-49).

Is it true that model-based allows behavioural flexibility?

Our comments about the utility of model-based and model-free RL were aiming to outline the computational properties of these different classes of algorithm, rather than their behavioural consequences. We apologize if this was not clear, we have modified the text to ensure we correctly cite the relevant literature when we make these statements. Regarding computational properties, we think it is an accurate characterisation to say that model-based RL algorithms can adapt to changes in the environment faster than a model-free RL algorithm. The reason for this is that a local change in the environment can change the optimal policy globally – e.g. if the rewarded goal location in a maze changes, this may require different decisions at locations far removed from the goal. When a model-based algorithm which has learned a model of the environment observes a local change in either the reward or transition structure, it can calculate the consequences of this for policy across the entire environment, albeit at the computational cost of planning. A model-free algorithm by contrast must learn the new correct policy via experience.

If this is the case, MB learning should be higher in the changing environment compared to the fixed version case because more flexibility is needed in the former; Is there any evidence this true?

In accordance with the above discussion regarding block lengths being similar in the two versions of the task, we expect the fixed and changing tasks should require a similar level of behavioural flexibility, given that the correct choice changes equally often in both. The only difference is that in the fixed task this is always because the reward probabilities have changed, whereas in the changing task this can be because the reward probabilities have changed, or because the transition probabilities have changed. It results that the adaptation of model-based control is spread slightly differently between the two tasks. However, as discussed above, from the perspective of a model-free RL algorithm these tasks look very similar, and a majority of subjects are model-free pre-debriefing.

We also note that previous studies have argued for arbitration between model-based and model-free control of behaviour on the basis of the uncertainty in each system about the best policy (Daw et al. 2005 Nature neuroscience 8 (12), 1704-1711, Lee et al. 2014, Neuron 81 (3), 687-699). The non-stationary transition probabilities in the Changing task would be expected to increase uncertainty about the transition probabilities used by the model-based system, which according to this account would be expected to reduce the influence of model-based control on behaviour. Consistent with this, in the Changing task we did not see the increased loading on the transition-outcome interaction predictor with experience that we observed in the Fixed task (Figure 2D, S10B), and removing the model-based component of the RL model had less impact on the fit quality in the Changing task than in the Fixed task (Figure S2A, S10D).

This is now discussed in greater detail as:

Supplementary information, lines 24-28: *These data suggest that, while model-free RL was dominant for the non-instructed sessions of both tasks, the dynamically changing action-state transition probabilities in the Changing task further reduced the ability to learn a model-based strategy. This is consistent with uncertainty-based arbitration between model-based and model-free control*[33,4]*, as the changing transition probabilities would be expected to increase uncertainty in the model-based system.*

I addition, authors mention that model-free learning allows rapid action selection but uses information less efficiently. Is there any support for this in the current experiment? If anything, in MF learning both action values can be updated on each trial whereas in MB learning only value related to the observed transition can be updated, making MF faster and more efficiently. All these claims can be tested in the current study by comparing learning rates and weight of MB vs, MF learning between the Fixed and Changing versions.

We appreciate the question regarding efficient use of information with MF vs. MB learning. Indeed, our data is consistent with model-based RL using information more efficiently, since loading on the transition-outcome interaction predictor across sessions 1 to 3 (a measure of model-based RL) was positively correlated with the number of rewards obtained by each subject, for both tasks. This is reported in the main manuscript, and has been revised for clarity, as:

Results, lines 163-167: *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=0.41, P<0.001, Pearson's correlation), suggesting that subjects who learned a model of the task used information more efficiently and thus obtained rewards at a higher rate.*

This finding is also consistent with current consensus that, in standard model-free RL, only the value of the chosen action is updated at each step, which in the two-step task corresponds to only one of the first-step actions being updated on each trial. By contrast, for a model-based agent, the change in value of the second-step state due to the trial outcome will update the value of both first-step actions according to the subjects' current estimate of how likely they are to transition to that state.

Regarding comparisons between the two tasks, as discussed above, prior to receiving explicit information about task structure, the predominant strategy in both tasks was model-free RL, with the fitted parameters further indicating that trial outcomes (rewarded or not) primarily directly reinforced the preceding first step choice, irrespective of the particular second-step in which the reward was obtained. This is consistent with the observed large main effect of outcome on repeating choice, as observed in the logistic regression. From the perspective of such a 'direct-reinforcement' model-free strategy, which is insensitive to where the rewards are received, the fixed and changing tasks are very similar. Consistent with this we did not see any significant differences in the RL model fits between the fixed and changing task in sessions 1 or 3, and the proportion of participants developing model-based RL at session 3 was similar across the two versions of the task (~14-15%). In any case, as discussed above, the logistic regression and model-comparison analyses provide some evidence for some, although limited, use of model-based RL in the Fixed task, consistent with uncertainty-based arbitration between strategies. Equivalent evidence was not found for the Changing task. Similarities and differences between the two tasks in the uninstructed phase are now discussed in greater detail in the Supplementary information:

Supplementary information, lines 8-9: *The proportion of subjects for whom a likelihood ratio test indicated model-based RL was being used in session 3 (6/42) was similar to that observed in the Fixed task (10/67).*

Supplementary information, lines 24-28: *These data suggest that, while model-free RL was dominant for the non-instructed sessions of both tasks, the dynamically changing action-state transition probabilities in the Changing task further reduced the ability to learn a model-based strategy. This is consistent with uncertainty-based arbitration between model-based and model-free control33,4, as the changing transition probabilities would be expected to increase uncertainty in the model-based system.*

-- One of the main points of the current study is that MF learning is more prominent and adopted first. Two recent studies using similarly complex task with multidimensional stimuli (and with no explicit instructions) have shown that feature-based learning, which resembles model-based is the starting learning strategy before transitioning to more accurate object-based learning (Farashahi et al, Nat Comm 2017, Cognition 2020). How do authors see the results of these studies reconcile with their findings?

Feature- vs object-based learning is interesting, and we appreciate the opportunity to discuss our results in the context of the work by Farashahi and colleagues. Our understanding is that the distinction between feature-based and object-based learning is, at least in principle, largely orthogonal to that between model-based and model-free learning. The model-based vs model-free dichotomy concerns what subjects learn to predict about the future. Specifically, whether individuals learn to predict future reward directly (model-free), or to learn to predict future states and likely contingent outcomes, and hence estimate expected future reward indirectly (model-based). The feature- vs object-based distinction by contrast, is to do with whether and how subjects generalise learning among items based on their sensory features. Indeed, one could imagine, for future research, that the task described here could be enriched with stimuli that have multiple feature dimensions, only some of which are relevant. Given our observation that instructions appear to shape subjects' internal representation of which states are important or distinct, even in a task where states were explicitly signalled, it would be very interesting to understand instruction effects in the more ambiguous conditions investigated by Farashahi and colleagues, where relevant features must be inferred.

We discuss this in the revised manuscript as:

Discussion, lines 594-597: *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}.

-- The main results are quite similar across healthy subjects and individual with OCD and other mood disorders (with some minor differences.

Does this mean that this task (and the original task) are not very useful to study changes in MB vs. MF learning in the clinical population, or this dichotomy is not the best way to look at reinforcement learning (see Collins and Cockburn, Nat Rev Neuroscience 2020)?

As discussed above response to an earlier question from reviewer 2, we believe that both the original Daw two-step task and the uninstructed version presented here provide valuable and complementary information about human behaviour. They examine distinct situations, both of which are common in real-life decision making. Specifically, our data speaks to situations where subjects must learn from experience to act in an unfamiliar domain, and show that behaviour in these circumstances is very different to that in situations where subjects are given detailed information about task structure. Also as discussed previously, we agree that the limitations pointed out by Collins and Cockburn paper are very much worth discussing and this has been added to the discussion section of the revised manuscript as:

Discussion, lines 584-590: *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*^{12,33,63}*, and can obscure their dependence on common computational primitives such as a representation of the task state-space*^{64}*. Although standard model-free and model-based algorithms provided a better fit to subjects 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.*

We don't think our findings of limited differences between OCD patients and healthy volunteers in their use of model-based RL invalidate the previous findings with the Daw task, among other reasons because the two tasks are distinct, and we did not compare them directly. Rather, we propose that our data suggest that previously unexplored aspects of how tasks are presented can affect the extent and nature of clinical differences, and have proposed potential clinical implications of such findings in the discussion.

-- Was there any significance difference in the learning rates between fixed version and changing version of the task. I ask this because there is an intuition that volatility should increase learning rates, and although a study by Behrens et al (Nat Neuroscience 2007) has provided evidence for such an intuition Farashahi et al, Nat Hum Behaviour 2019 found no evidence for it.

As discussed above, when the subjects are learning from experience, the block lengths in the Fixed and Changing task are very similar, so the volatility of the tasks is matched with regards to the predominantly model-free strategy the subjects employed. We therefore don't think our data speak to the question of whether/how environment volatility affects learning rates. Consistent with this, as we noted in response to an earlier comment from this reviewer, we did not observe significant differences in RL model-fits to the fixed and changing task pre-debriefing.

-- Was there any significance difference in the learning rates between control and clinical groups?

We do not see any significant differences in the effects of either uninstructed experience or debriefing on learning rates between either clinical group and healthy volunteers. Significant effects are reported in the main text and we now report the full set of permutation test results in new supplementary tables S1 and S2 (shown below).

Supplementary table S1 – Permutation test results for differences in learning and debriefing effects between healthy volunteers and individuals with OCD.

Parameter	**Learning effect**			**Debriefing effect**		
	Group diff. in parameter change	95% CI of null hypothesis test statistic	P value	Group diff. in parameter change	95% CI of null hypothesis test statistic	P value
Logistic regression						

Outcome	0.28	-0.464, 0.477	0.25	-0.36	-0.528, 0.537	0.18
Transition	-0.07	-0.299, 0.295	0.65	-0.20	-0.489, 0.485	0.44
Trans. outcome	-0.43	-0.350, 0.359	**0.02**	-0.20	-0.584, 0.597	0.54
RL model						
Model-free strength (MF)	1.52	-1.681, 1.684	0.07	-0.04	-1.615, 1.568	0.94
Model-based strength (MB)	-0.27	-0.579, 0.574	0.37	0.04	-0.831, 0.834	0.94
Value learning rate (αQ)	0.01	-0.282, 0.265	0.91	-0.08	-0.253, 0.262	0.54
Eligibility trace (λ)	0.05	-0.204, 0.229	0.67	-0.01	-0.198, 0.203	0.93
Transition learning rate (αT)	0.05	-0.411, 0.374	0.75	0.02	0.315, 0.308	0.91
Choice bias	-0.08	-0.299, 0.299	0.62	0.24	-0.355, 0.354	0.18

Choice perseveration	0.22	-0.577, 0.572	0.48	-0.86	-0.842, 0.800	**0.04**

Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between healthy volunteers (n=67) and individual with OCD (n=46) in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3).

Supplementary table S2 – Permutation test results for differences in learning and debriefing effects between healthy volunteers and individuals with mood and anxiety disorders

Parameter	**Learning effect**			**Debriefing effect**		
	Group diff. in parameter change	95% CI of null hypothesis test statistic	P value	Group diff. in parameter change	95% CI of null hypothesis test statistic	P value
Logistic regression						
Outcome	0.32	-0.441, 0.437	0.16	-0.15	-0.487, 0.507	0.57

Transition	0.06	-0.291, 0.291	0.68	-0.10	-0.507, 0.518	0.68
Trans. outcome	-0.19	-0.378, 0.387	0.35	0.13	-0.594, 0.617	0.68
RL model						
Model-free strength (MF)	0.27	-1.342, 1.262	0.67	0.22	-1.229, 1.241	0.75
Model-based strength (MB)	0.06	-0.550, 0.550	0.87	0.49	-0.750, 0.732	0.20
Value learning rate (αQ)	0.08	-0.261, 0.253	0.53	-0.08	-0.254, 0.250	0.53
Eligibility trace (λ)	-0.08	-0.170, 0.198	0.37	0.08	-0.200, 0.191	0.43
Transition learning rate (αT)	-0.16	-0.556, 0.516	0.65	-0.12	-0.439, 0.443	0.60
Choice bias	0.11	-0.273, 0.271	0.42	0.11	-0.336, 0.328	0.51

Choice perseveration	0.50	-0.598, 0.609	0.11	-1.40	-0.821, 0.840	**0.0012**

Permutation tests (5000 permutations) were used to assess differences in the fitted model parameter loadings between healthy volunteers (n=67) and individual with mood and anxiety disorders (n=49) in the effect of learning (defined as change between session 1 and 3) and debriefing (defined as change between session 3 and 4, taking only subjects who are MF at session 3).

-- In some places, authors rely in 0.05 as the threshold for statistical significance. Considering the number of comparisons (e.g., between model parameters) I don't think 0.05 is an appropriate threshold.

Our manuscript contains both important positive results and important negative results, i.e. in several places it is the absence of a significant effect that is striking given the prior literature. We therefore think it is important to balance the risk of false positives against that of false negatives, as we do not want readers to come away with the impression that where we did not observe differences where they might be expected, this was because we had a high risk of false negatives due to multiple comparison correction. Our approach in the manuscript has therefore been to report uncorrected P values in full (as opposed to as P<0.05 etc), such that readers can directly evaluate the strength of the statistical evidence. It is worth noting that most of our positive results have very robust P values that would survive multiple comparison correction for multiple model-parameters. We also note that we replicated our key findings concerning behaviour of healthy volunteers to a striking extent in the new dataset gathered with a slow paced version of the task to address reviewer 1's comments (figure S4 & S5). Specifically, we confirmed that initial behaviour is model-free, that model-based increases in a minority of subjects with experience, and that subsequently providing explicit knowledge strongly boosts use of model-based but also affects model-free value updates.

-- Why there is no comparison between stay probability in different conditions (to prove the unbalanced pattern of stay probability as a function of outcome and transition)?

Given previous work from some of the authors of this paper, we think that reporting statistics based on the logistic regression analysis is the most principled way of quantifying the influence of trial events on the subsequent choice. Specifically, we have used the logistic regression analysis (e.g. Figure 2D) to quantify significant effects of outcome, transition and their interaction on stay probability. This analysis models the same data as shown in the raw stay probability plots, but importantly allows us to incorporate other influences on subjects' choices such as choice biases. Akam et al. (PLOS Comp. Biol. 2015) have previously shown that in two-step tasks with a strong contrast between good and bad options (including the one used here), correlations across trials can cause even model-free agents to show a significant influence of transition-outcome interaction on stay probability, because the difference between chosen and non-chosen action values at the start of the trial is correlated with the subsequent transition-outcome interaction. In that paper it is shown that this can be corrected for by including an additional predictor in the logistic regression analysis of stay probabilities, allowing unbiased estimation of the effects of the trial events on the next trials stay probabilities. We have clarified this approach in the methods section of the revised manuscript as:

Methods, lines 721-724: *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 subjects biases and cross trial correlations (see below) to be taken into account.*

Methods, lines 729-732: *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]*.*

-- It seems to be a general increase in stay probability after debriefing as a side effect. What could be a reason for this?

The reviewer is correct that stay probabilities overall increase following debriefing, and this also shows up in the perseveration parameter of the RL model, which increases following debriefing. We think this occurs because during the debriefing subjects are told that the reward probabilities at the two sides reverse only occasionally, and therefore, once they have detected a switch they expect the reward probabilities to be stable for an extended period. To examine this explicitly, we analysed whether the increase in perseveration as a whole due to debriefing correlated with a reduction in perseveration over the course of each block in the post debriefing data (Figure S7, right panels). The rationale is that if the increase in perseveration due to debriefing is due to expecting a period of stable reward probabilities after a block transition, then in subjects who show this effect strongly, we should also expect to see that, after debriefing, they are actually less perseverative late in blocks compared to early (whilst being more perseverative overall). This was indeed the case in healthy volunteers, who showed the debriefing effect on perseveration strongly, but not in either clinical group, both of which showed a significantly weaker (OCD group) or non-statistically significant (mood and anxiety group) effect of debriefing on perseveration. This analysis is discussed in the manuscript as:

Results, lines 341-351*:* *Debriefing also increased how often subjects repeated choices independent of subsequent trial events, as reflected by a significant increase in the 'perseveration' parameter of the RL model (95% CI of null hypothesis test statistic [-0.75,0.76], parameter change=1.63, P<0.001; 95% CI of null hypothesis test statistic [-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=-0.35, P=0.02; Pearson's correlation; Figure S7).*

-- Authors focus on participants who did not use MB learning in session 3 (e.g. in Figure 4), but what did happen to participants who adopted MB after they received instruction?

It is an interesting question to understand what happens to behaviour on a task subjects have already learned a good model of from experience, when they are subsequently told the structure. Unfortunately, our data does not really allow us to speak clearly to this, because in the debriefing group on the fixed task, only 3 subjects had adopted model-based at session 3 as assessed by the likelihood ratio test. In these 3 subjects we do not see any obvious changes in behaviour as a result of debriefing, but we feel the sample size is so small that it is not worth reporting this in the manuscript. This analysis is shown below for the reviewers' interest:

<u>Debriefing effects in the 3 subjects on the fixed task who were model-based at session 3.</u>

Minor

-- Figure 4 caption: "in control"

-- Figure 5 caption: "in OCD and other mood disorders?"

We have corrected figure captions to clarify the target population for the data in each figure. We opted not to include ‘and other mood disorders’ given that this specific population was mostly intended as a control for the OCD population, both regarding the presence of anxiety/depression symptoms and the effects of medication.

Reviewer #3:

Remarks to the Author:

In the present manuscript, Castro-Rodrigues and colleagues used a modified version of the ‘two-step task’ task to assess contributions of model-based (MB) versus model-free (MF) systems to reinforcement learning in humans. Specifically, they tested to what degree humans deploy MB vs MF learning, first in the absence of any explicit instruction about the task structure, and then after fully explicit debriefing. Both healthy humans and individuals suffering from OCD (and another clinical control sample) were tested. The key finding according to the authors is that uninstructed behaviour was model-free, with model-based control only emerging over time in a subset of participants. The latter was less pronounced in the OCD group. All groups showed stronger model-based behaviour after receiving explicit debriefing on underlying task structure.

The manuscript is well written and addresses an important and interesting question, particularly given influential earlier reports that increased MF control may be a common feature of compulsive behavioural disorders (Voon et al., 2014). However, I am concerned whether the conclusions drawn by the authors are justified by the data and the study design.

We thank the reviewer for the recognition of the interest of the questions addressed. We have added several additional analyses to address the reviewers’ concerns, as detailed in response to individual comments below.

1) One of the main findings is that uninstructed behaviour was dominantly model-free. This is a strong statement given the results and the analyses presented. In my view, what can be said with confidence is that subjects did not use the *true* model of the task - and this is the only test of 'model-based behaviour' the authors are presenting. Subjects could have used a completely different model of the task which would not be evident from testing to what degree they used the optimal model. Or indeed volunteers could have tried out different models of the task throughout the three sessions. For instance, as you present in Fig. S1, quite a proportion of people keep pressing invalid keys for a considerable period throughout at least the first session. This implies that it took subjects fairly long to understand even the most fundamental task characteristics. The authors write about there being no evidence of participants using task structure early on (line 337/338) - but given the above, everything else would be highly surprising!

Indeed, da Silva and Hare (which is also cited here) have shown that providing subjects with inaccurate models of the task evokes MB behaviour that can appear completely model-free. They also describe how behaviour which appears to be a hybrid of model-based and model-free could also be explained by a set of different algorithms (see also Collins & Cockburn, 2020). In this context, it would be important to see model comparisons that include competing models as well as model simulations on the eventual winning model.

This is an important issue, and we thank the reviewer for bringing it up and allowing us to address it in more detail. Our understanding of Silva and Hare's data is that they show that agents which are completely model-based, but have an incorrect model, can generate behaviour that looks similar to an agent that uses a mixture of model-based and model-free. As far as we can tell, all of their simulations from model-based agents with incorrect models still show a substantial transition-outcome interaction (the classical signature of model-based) but also show a significant main effect of outcome (normally considered a sign of model-free learning). This pattern is very different from that of our subjects at session 1, where we see a large main effect of outcome, but no influence at all of the transition outcome interaction. This is precisely the pattern expected for an essentially pure model-free strategy in which outcomes directly reinforce the preceding choice, and we are not aware of any incorrect-model-based strategies which generate behaviour that looks like this.

Nonetheless, we agree with the reviewer that, given the Silva and Hare data, it is important to explicitly test whether the types of incorrect-model-based strategies they consider might in fact explain our subjects' data. We therefore did an additional model-comparison on data from session 1 and session 3, in which we included the two incorrect-model agents proposed by Silva and Hare (termed 'unlucky symbol' and 'transition-dependent learning rate'), as well as the incorrect-model proposed below by the reviewer. Model-comparison indicated that all 3 of these incorrect-model agents fit data from both session 1 and session 3 less well than any of the traditional models we had previously considered. According to the reviewer's suggestion, we simulated behavioural performance of an agent using the best fitting model (mixture model) and observed it produced stay probability plots which were qualitatively similar to the collected data (Figure S6).

These results are presented in figure S2B and figure S6 (reproduced below), and discussed in the revised manuscript as:

Results, lines 178-183: *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 structure32, 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 (Figure S2b). We also simulated behaviour from the best fitting RL model and verified that it produced stay probability plots qualitatively similar to the experimental data (Figure S6).*

Supplementary figure S2b) Model comparisons including additional model-based agents with incorrect models of the task structure; one which believed state transitions were deterministic but volatile (IM-DV), one with transition dependent learning rates at the second-step (IM-TDLR) and one which believed that one first step option was unlucky and reduced reward probability at the second-step (IM-US).

Supplementary figure S6. Stay probabilities for best fitting RL model 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). Top panels show experimental data, bottom panels show behaviour simulated from the best fitting RL model, which used a mixture of model-based and model-free. Error bars indicated the cross subject standard error of the mean (SEM). In each group data was analysed separately for session 1 (blue graph), session 3 (red graph) and session 4 (gold graph).

2) An alternative model employed by participants could be that, rather than assuming a probabilistic, but fixed mapping of first-step choices to second-step states, participants might assume this mapping to be fully deterministic, but highly volatile. Under such a model, following a rare transition, participants would infer that the underlying latent state has changed and assume that their first-step choice would now lead them to the state 2 they just observed. An agent using such a model would appear exactly like a completely MF agent. This could also be a potential explanation for the effects seen in the OCD group in figure 3 (A and D). It shows that, in OCD, there is an increase in MF control

(Gmf) from session 1 to session 3 - and likewise an increase in the effect of outcomes. Again, I am not convinced that such a pattern is fully indicative of MF behaviour. As discussed above, the increase in P(stay) following rewards (irrespective of transition type) could as well be obtained from an inaccurate world model. Thus, such a pattern may actually indicate *increased* MB control in OCD!

We thank the reviewer for this suggestion. If we understand it correctly, the proposed agent is actually a special case of our model-based agent, because our model-based agent learns the transitions from experience, and if its learning rate for transitions (parameter) is set to 1, the agent believes that the most recently observed transition for a given action occurs with probability 1 when that action is selected. We would not expect this agent to generate a pattern of stay probabilities similar to that observed in our session 1 subjects, because model-based agents with a high learning rate for transitions generate strong loading on the transition predictor, in addition to the outcome predictor (Akam et al. PLOS CB 2015, Figure S5). Nonetheless we included it in the new model-comparison discussed above (figure S2B) but found that, as with the other incorrect-model agents, it fit the data worse than any of the traditional models that were considered.

3) I appreciate the motivation to simplify the original two-step task. However, by removing the choice alternatives at the second-step state, the task, at least to subjects, may lose its key characteristics as a multistep decision problem, but instead be perceived as a one-stage decision problem with probabilistic rewards (see my comment above). This might have influenced the perceived importance of forming a model and the consideration of the sequential structure of the task. Additionally, as the authors also note, the new design greatly reduces working memory load. Relationships between working memory capacity and task complexity on one hand and decision strategy on the other have been reported previously (Otto et al., 2013; Kim et al., 2019), which may also contribute to the apparent absence of MB behaviour.

Our motivation for removing the choice at the second-step was, as the reviewer points out, to simplify the task as much as possible, as we felt that the more complex the task was, the harder it would be for subjects to learn and use a model. Nevertheless, we appreciate the concerns raised by the reviewer, and have added to the discussion regarding the possibility that behaviour could be different in tasks with larger state-spaces:

Discussion, lines 590-594: *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.*

4) Furthermore, it was not clear to me why the authors opted for a fixed spatial-motor mapping instead of presenting two distinct visual stimuli with random mapping to top/bottom position on the screen. As the authors acknowledge in the discussion, such a fixed mapping of spatial position to effector may have further encouraged the use of a habitual/model-free/S-R strategy.

Again, the motivation was to simplify the task as much as possible, as we expected this would make it easier for subjects to learn a model of the task. As the reviewer notes we do discuss the possibility in the discussion that using fixed spatial-motor contingencies may affect the strategy used. We think it is equally possible that using discriminative stimuli whose position is randomised could make it harder, rather than easier, to use model-based RL, because action-outcome predictions and stimulus-outcome predictions, thought to be mediated by at least partially separate brain systems (Rudebeck et al. J.Neurosci 28.51 (2008): 13775-13785), would no-longer be aligned.

5) Modelling:

a) it would be re-assuring to see whether the fitted parameters can be recovered from simulated data generated using the fitted parameters (parameter recovery).

We now include a new parameter recovery analysis in which we test how well individual parameters can be recovered when we set the other parameters to their fitted values from either session 1, 3 or 4, and use the same number of simulated subjects as we have real subjects (Supplementary figure S8, reproduced below). This is discussed in the revised manuscript as:

Results, lines 352-356: *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 (Figure S8). 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.*

Supplementary figure S8. RL model parameter recovery. Test of the accuracy with which RL model parameters could be recovered from simulated data. Panels show mean and standard deviation of recovered parameters across 10 repeated simulation runs when the parameter under investigation was fixed at the specified 'true parameter value' and the other parameters were drawn randomly for each subject from the population level distributions fit to the specified dataset (top row- fixed task session 1, middle row – fixed task session 3, bottom row – fixed task debriefing group session 4).

b) Unlike in Daw 2011, there are separate weights (Gmb, Gmf) for the contributions of the MB and MF systems, respectively (rather than Gmf = 1 – Gmb). This seems plausible to me, as there may well be participants in which both MB and MF is low (e.g. purely stochastic or perseverative choice), but since this is also a - minor - departure from the original analyses, it would be helpful to briefly justify this in the text.

We thank the reviewer for this suggestion. We included a justification for the separate weights in the methods section of the revised manuscript (lines 769-771).

Methods, lines 783-785: *We used separate weights (G_mf, G_mb) for the influence of the model-based and model-free systems[30], 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].*

c) In figure 2F, the modelling results do not support the logistic regression - there is no difference in Gmf and Gmb between session 1 and 3.

The reviewer is correct that there is a discrepancy between the increase in the influence of the transition-outcome interaction on subsequent choice in the logistic regression between session 1 and 3 and the absence of a significant change in the model-based weight parameter of the RL model. This is discussed in the manuscript (lines 189-194) where we state:

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 subjects learned to use model-based RL; as model comparison for individual subjects between the mixture RL model and a simpler model-free RL model, indicated that only 15% of subjects (10/67) used model-based RL at session 3 (likelihood ratio test, threshold P=0.05).

We note that the maximum-a-posteriori fits for individual subjects (dots in figure 2F) indicate that a small number of subjects have approximately double the influence of model-based RL at session 3 compared to session 1, but this does not drive a significant change at the level of the population as a great majority of subjects show minimal change.

We also note that in addition to the model-based weight parameter, the value learning rate will also affect the transition-outcome interaction as it determines how much trial outcomes update the value of second-step states, which are used by the model-based system to compute the value of first step actions. This parameter did show a significant increase between session 1 and 3.

d) Bias parameter in the model: is my understanding correct that this indicates an action bias for the top circle? If so, why is it B = 1 for the high and B = 0 for the low action? Would this not mean that the model can have a bias for the upper circle, or no bias at all? Shouldn't B = –1 for the low option to also allow for a bias toward the lower option?

Apologies, this was not explained clearly enough in the methods. We have modified the relevant section to clarity how the bias and perseveration parameters work, it now reads (lines 764-773):

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

where and are parameters controlling, respectively, the strength of influence of model-free and model-based action values on choice. is a parameter controlling the strength and direction of choice bias, is a variable which takes a value of 1 for the up action and 0 for the down action. Since it is the difference in values between up and down actions that determines choice, positive values of therefore generate a bias towards the up action and negative values towards the down action. is a parameter controlling the strength and direction of choice perseveration, is a variable which takes a value of 1 if action was chosen on the previous trial and 0 if it was not. Positive values of therefore promote repeating the previous choice while negative values promote switching.

e) Decreases in the eligibility trace parameter lambda correlate, across subjects, with increase in MB control - is this naturally arising because as lambda --> 0, there is no more update in the MF system? In other words, is it possible, for a given subject, that there are two (probably very similar) local optima in the parameter space, at low (high) values for lambda and high (low) values for Gmb?

The eligibility trace parameter does not affect the relative influence of model-based and model-free values on choices, it just changes the way that model-free values for first-step actions are updated. Specifically, when is 0 the update to first-step action values depends only the value of the second-step reached, and when is 1 the update depends only on the trial outcome, with intermediate values of determining the mixture of these two influences on the update. We therefore would not expect problems in accurately fitting as long as there is a model-free influence on choices (as only affects updates to the model-free first step action values). We verified this empirically with the new parameter recovery analysis (figure S8, discussed above), confirming that we are able to accurately recover the values of both the model-free weight and when the other parameters of the simulated data match those fitted to the real data from either session 1,3 or 4.

6) In figure 2F, the modelling results indicate no difference between MB and MF contributions to behaviour. This is at odds with the logistic regression results presented in 2D (and the p(stay) in 2C) which indicates a clear transition X outcome interaction. This again raises the issue to what extent the model with the fitted parameters is able to recapitulate the actual pattern in subjects' choice behaviour (related to my question above regarding parameter recovery)

If we understood correctly, we think the reviewer might be reiterating the point here that they made in comments 5a and 5c. We now included a parameter recovery analysis and further discussed these issues in lines 346-350 and lines 186-191, as well as in our above responses to comments 5a and 5c.

7) In the analyses investigating the effects of providing the full task structure, it seems that out of the 57 healthy individuals not showing MB behaviour yet, n = 41 were assigned to the debriefing group whereas only n = 16 were not debriefed. Why were they evenly allocated to both groups? Comparing the proportion of subjects acquiring MB control appears odd given the highly unbalanced group size?

Initial experiments were conducted among healthy volunteers recruited in Lisbon, with two variants of the task (Fixed and Changing task). 82 participants were randomized between the two versions of the task and, within each task, some participants were debriefed between sessions 3 and 4 (17 performing the Fixed version and 16 performing the Changing version of the task), while the remaining subjects were not debriefed (23 in the Fixed version and 26 in the changing version). From the 23 non-debriefed subjects in the Fixed version, 16 were not showing model-based at session 3. Given the results from this first sample, in healthy volunteers subsequently recruited in New York (n=27), as well as all clinical samples in both sites, only the Fixed task was used, and all participants were debriefed. Thus, the n=41 subjects who were not model-based at session 3 were 25 of the 27 healthy volunteers recruited in NY and 16 of the 17 healthy volunteers who performed the Fixed task with debriefing in Lisbon. As a result, the sample is unbalanced towards the participants performing the Fixed task and that were debriefed. To address potential problems resulting from the unbalanced nature of the debriefed and non-debriefed samples, as mentioned in the results section, analyses were repeated including only participants recruited in Lisbon, which did not affect the results.

Furthermore, we tested differences in learning or debriefing effects between the Lisbon and New York debriefing groups, and did not find any significant differences (Table S3).

This is now clarified in the methods section, lines 698-701, where it now reads: *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.*

8) Why are people diagnosed with a wide variety of mood and anxiety disorders put together as one group? These disorders are characterized by several different symptoms which could also be expected to influence MB/MF control of behaviour in different ways. I guess it also does not serve to "investigate potential contributions of medication or unspecified mood and anxiety symptoms" for the same reasons. Further details on psychotropic medication would be needed to evaluate their comparability with respect to medication.

We decided to include a group of participants with different mood and anxiety disorders because there is a very high comorbidity between OCD and mood disorders (such as depression) and anxiety disorders (such as generalized anxiety disorder). Also, to the best of our knowledge, there is no consistent evidence for imbalance in model-based/model-free control in any specific mood or anxiety disorder. Nevertheless, the choice of a group of several disorders as a 'control', rather than a specific disorder, reduces the risk of observing effects dependent on the 'control' condition, rather than effects dependent on the disorder of interest (OCD). Importantly, scores in assessments of anxiety and depression symptom severity were equivalent between the two clinical groups, that were only different according to severity of OCD symptoms. This suggests that this group can in fact allow for control of the effects of unspecific mood and anxiety disorders, while studying the more specific effects of OCD symptoms.

Regarding medication, we classified it in classes (SSRI, tricyclic antidepressant, second-generation antipsychotic, first-generation antipsychotic, mood stabilizers, benzodiazepines, other antidepressants and other medications) and we did not find differences between groups in the use of any class of medication (P's>0.2; Chi-squared or Fisher's exact test) (see Table below). It is also important to mention that while the clinical groups recruited in Lisbon were medicated, the clinical groups recruited in New York were unmedicated. As reported in the original manuscript, we did not find differences between Lisbon and New York groups in any behavioural or psychometric measure.

Table r1 – Psychotropic medication in each group. Data was compared between groups using Chi-square test or Fisher's exact test (when samples were very small). OCD = Obsessive-compulsive disorder; MA = Mood and anxiety disorders group; SSRI = Selective serotonine reuptake inhibitor; TCA = Tricyclic antidepressant; SGA = Second-generation antipsychotic; FGA = First-generation antipsychotic; BZD = Benzodiazepine).

We now present this in the manuscript, lines 113-115, where it now reads: *Regarding medication, we classified it in classes and we did not find statistically significant differences between groups in the use of any class of medication (χ2<1.6; P's>0.2; across all Chi-squared tests).*

9) RT at 2nd stage are faster in common compared to rare transitions, and this effect is present already in session 1 (but seems to become larger in session 3). Is this an indication that subjects have learnt the state transition probabilities in session 1 already?

The reviewer is correct that reaction times at the second-step are faster following common than rare transitions even in the first session where choices are model-free. We think this likely reflects motor systems learning to predict, and hence prepare, upcoming actions before higher-level decision-making systems have learnt to use state predictions to guide choices. This is discussed in the manuscript as:

Results lines 148-153: *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.*

Results lines 195-205: *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, Figure S1). 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 subjects with the rare-common reaction time difference at both session 1 (r=0.57; P<0.001; Pearson's correlation, figure S3) and session 3 (r=0.69; P<0.001), suggesting interaction between learning at motor and cognitive-levels, though this appeared to be driven by the minority of subjects whose choices were more model-based.*

See also our response to reviewer 1's question about this.

10) Before debriefing participants, did you question them about what kind of strategy they used? While this would probably hard to quantify, it might at least give some qualitative hints as to what beliefs they had acquired (and what models of state space they applied) during the 900 uninstructed trials.

We thank the reviewer for this question. Prior to informing subjects about the task structure we did indeed assess their beliefs, using a pen-and-paper questionnaire given at the end of session 3. We had not included this data in the previous version of the manuscript because we were unable to access the documents with the NY groups responses due to COVID-19 restrictions preventing access to the building. We have now been able to analyse this data for all subjects and include the results in new Supplementary information (lines 50-102).

Participants were required to answer four questions. For the first two questions, they were asked to provide open-ended answers. Afterwards, the same questions were asked in a multiple-choice format.

The questions were asked after showing the participants the following image in a sheet of paper:

Specifically, the first question was: "If you press the upper arrow key when this screen is being shown, what is most likely to happen next?". The second question was: "If you press the lower arrow key when this screen is being shown, what is most likely to happen next?"

After the participants wrote their answers, the researcher collected the first sheet of paper and provided them a second sheet of paper, which also contained the same image at the top and the same two questions. However, instead of writing an open answer, participants were asked to choose one of four different options for each question:

a) The left side circle will turn yellow

b) The right side circle will turn yellow

c) A coin will appear on the left side

d) A coin will appear on the right side

To classify the open answers that participants gave as correct or incorrect, we created a set of criteria that the answers had to fulfil in order to be considered correct. According to these criteria, each answer had to contain at least one of the following elements: a) “the right circle would turn yellow / light up / be highlighted”; b) “the yellow circle moves [from the top circle] to the right circle”); c) “most of the times, the right circle would turn yellow, although in a minority of times the left circle turns yellow”. Two independent raters (PCR & AM) assessed open answers independently. Then, when in disagreement, they discussed and reached a consensus if those answers should be considered correct or incorrect. The rate of concordance between independent raters was 92%.

Considering all participants together, we found no statistically significant association between correct or incorrect answers and pre-debriefing behavioural measures, either in open answers (-1.7<t’s<0.2; P’s>0.09; independent sample t-test’s) or in multiple-choice questions (-1.2<t’s<0.1; P’s>0.2). However, we found that subjects who gave correct multiple-choice answers had a higher influence of model-based action values on choice after the debriefing (t = -2.66; P=0.009; independent sample t-test).

Concerning open answers that did not fill the criteria for being considered correct, we identified a frequent type of wrong belief/model, specifically that the second-step circle is random (~25% of open answers). Regarding the remaining incorrect open answers, they consisted of different types of answers each occurring with a small frequency (<10%), such as ignorance of the second step (“immediately after the first-step choice, a coin may appear or not”) or an incorrect transition model (common transitions identified as rare & rare transitions identified as common). As the first type of wrong model (random second-step) was particularly frequent, we divided subjects in three groups (correct model, random second-step, other wrong models) and tested if their pre-debriefing behaviour was different. We did not find statistically significant differences between groups in terms of behavioural strategy, either in logistic regression or model fitting analyses (F’s<2.0; P’s>0.14; one-way ANOVA).

When comparing different clinical groups, we found that while individuals with OCD had a smaller proportion of correct open answers (12/39) than healthy volunteers (23/43), this difference was not statistically significant (χ2 = 2.1, P=0.1, chi-squared test). We also did not observe differences between OCD and healthy volunteers in their proportion of correct multiple-choice answers (χ2 =0.45, P=0.5). Individuals with mood and anxiety disorders had the same proportion of correct/incorrect answers that healthy volunteers, both in open answers (χ2 =0.2, P=0.7) and in multiple-choice (χ2 =0.03, P=0.9).

Our findings complement previous data provided by other groups using an operant conditioning outcome devaluation paradigm in which OCD subjects did not differ from healthy volunteers in their capacity to provide explicit assessment of the contingencies governing action and outcome, although their behaviour did not reflect these (Gillan et al., 2014; Gillan et al., 2015). However, in those studies outcomes were aversive and the experimental induction of habits in humans using operant conditioning paradigms has been questioned (De Wit et al., 2018; Robbins et al., 2019).

11) It is unclear to me why the neutral blocks existed? What was their purpose?

The original two-step task (Daw et al. Neuron, 2011) used reward probabilities which drift as random walks on the range 0.25 - 0.75, so most of the time they are pretty close to neutral. This has the advantage of reducing correlations across trials, which can otherwise complicate analysing how events on one trial affect subsequent choices (see Akam et al. PLOS Comp. Biol. 2015 for a detailed discussion). However, it also means there is very little contrast between good and bad options, and means that model-based RL does not yield a significantly higher reward rate than model-free (Akam et al. PLOS Comp. Biol. 2015, Kool et al. PLOS Comp. Biol 2016). We felt it was important that there was a significant contrast between good and bad options to promote task engagement and ensure that model-based RL yielded higher reward rates, hence having 80-20 non-neutral blocks which only transitioned once the subject started choosing the correct option, but incorporated neutral blocks as well to try and reduce cross trial correlations, at least to some extent.

Minor:

1) In the introduction (Line 79 onwards) you write: "However, no studies have explored behaviour on multi-step tasks in the absence of information about task structure, in either healthy or clinical populations" <-- I can think of one study (Gläscher et al. 2010, Neuron) that used Tolman-style latent learning in an abstract multi-step T-maze. I think this is also without explicit instruction and likewise looks at the contribution of MF and MB systems over time.

We thank the reviewer for pointing out this relevant prior work. It is true that subjects in Gläscher et al. learned the task structure from experience (during the latent learning phase) then subsequently used this knowledge in the rewarded phase to guide choices. We have modified the introduction to state that ‘few’ rather than ‘no’ studies have explored this question, and now discuss the Gläscher et al study explicitly in the discussion as:

Lines 597-602: *Another question is how information given to subjects 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, subject 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].*

2) In the discussion (#341): "When learning from experience, individuals with OCD were impaired in their use of model-based control and biased towards a model-free strategy." This makes it sound as if this was only true for OCD, but it likewise applies for healthy controls.

We have clarified this in the revised manuscript, lines 441-443, where it now reads: *"When learning from experience, individuals with OCD were impaired in their use of model-based control and were more biased towards a model-free strategy, as compared with healthy volunteers."*

3) Line 283/285: "Increased use of model-based RL after debriefing was confirmed by model fitting (Figure 5d), which showed increased influence of model-based action values on choice" <-- I guess this should be Figure 5e?

Thanks, fixed in revised manuscript.

Decision Letter, second revision:

22nd October 2021

Dear Professor Oliveira-Maia,

Thank you once again for your manuscript, entitled "Explicit knowledge of task structure is a primary determinant of human model-based action," and for your patience during the peer review process.

Your manuscript has now been evaluated by the same 3 reviewers as previously, whose comments are included at the end of this letter. Although the reviewers find your work to be improved, they also raise some important outstanding concerns. We remain interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these concerns in the form of a revised manuscript before we make a decision on publication.

In your revision, we ask you to address the issues raised by Reviewer #3, and highlight in particular the need for a satisfying response to their concern #1. Moreover, as you revise your manuscript, we ask you to review a number of aspects of the analysis. In particular, we emphasize that the difference between a significant effect and a non-significant effect is not in itself a significant difference - such a difference needs to be demonstrated, for example through an interaction contrast. It's therefore inappropriate to logically contrast significant and non-significant effects. Where you compare differences in behaviour between the groups, you must support these with appropriate analyses in all instances. Please also report full statistics in Figures 2 and 3, and add CIs and dfs to all reported correlations.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within four to eight weeks. We understand that the COVID-19 pandemic is causing significant disruption for many of our authors and reviewers. If you cannot send your revised manuscript within this time, please let us know - we will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,

Marike

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
The authors have done an excellent job addressing all my previous comments - in particular, they have gone to the length of collecting an additional 20 subjects to address my most substantial concern, which is that the pacing of the task was favouring a model-free strategy. It turned out that the results from this additional data collection supported the authors' original conclusions. I am happy to recommend publication.

Reviewer #2:
Remarks to the Author:
The authors have responded to all my concerns and comments. Indeed, the response was refreshingly clear and straightforward, and I appreciate the amount of work they put into the revision. I have no further concerns and would like to congratulate authors for executing such an interesting study.

Reviewer #3:
Remarks to the Author:
The authors have performed an exhaustive revision (and even included additional data from a new experiment). Most of my previous points have been fully addressed. I do have two remaining points though:

1) The authors have explained their aim to simplify the task, and I understand the importance of having 'good' and 'bad' options and the need to reduce correlations across trials. To me, it appears that the task can be perfectly solved by an MF agent that completely ignores the 2nd state and only tracks the rewards following their actions. Such an agent corresponds to the MF algorithm the authors have used, in the case where the eligibility parameter lambda is set to 1. Indeed, fitted values for

lambda are close to 1 (as the authors also acknowledge), which would speak to that. My understanding is that this coupling breaks in the neutral condition - thus MB contributions should be more readily revealed there. My suggestion would be to present the data from the neutral condition separately - I think it has the potential to change the interpretation of the data: if there is already hints of MB contributions to choice in the neutral condition (consistent with what implicitly seems to drive RT already in session 1), then providing subjects with explicit instructions would be amplifying the use of - an already existing - MB strategy. If, on the contrary, even in the neutral condition, no MB contribution was evident, then I would describe the effect of instruction more as breaking an acquired S-R habit.

2) Thank you for including the parameter recovery. I have a question on how the recovery was done. It is stated that ‘the parameter under investigation is fixed at the specified true parameter value but the other parameters are drawn randomly for each subject from the population level distributions’. Why are data not generated with the full combination of fitted parameters (e.g. all parameters fixed at their fitted value per participant and all parameters recovered from those artificial data)? In my understanding, this approach would allow an interpretation of the individual model parameters.

PS1 (Q5d) - I had confused b and Bi! The modified description makes things even more clear, thank you.

PS2 (Q6) - apologies, this was indeed just the same as the above comment, as you correctly pointed out!

Author Rebuttal, second revision:

REVIEWER COMMENTS:

Reviewer #1:

Remarks to the Author:

The authors have done an excellent job addressing all my previous comments - in particular, they have gone to the length of collecting an additional 20 subjects to address my most substantial concern, which is that the pacing of the task was favouring a model-free strategy. It turned out that the results from this additional data collection supported the authors' original conclusions. I am happy to recommend publication.

We thank the reviewer for their positive assessment of the revised manuscript and for supporting publication.

Reviewer #2:

Remarks to the Author:

The authors have responded to all my concerns and comments. Indeed, the response was refreshingly clear and straightforward, and I appreciate the amount of work they put into the revision. I have no further concerns and would like to congratulate authors for executing such an interesting study.

We thank the reviewer for their positive assessment of the revised manuscript and recognition of the work done to address reviewer comments.

Reviewer #3:

Remarks to the Author:

The authors have performed an exhaustive revision (and even included additional data from a new experiment). Most of my previous points have been fully addressed. I do have two remaining points though:

We thank the reviewer for their positive assessment of the previous revisions.

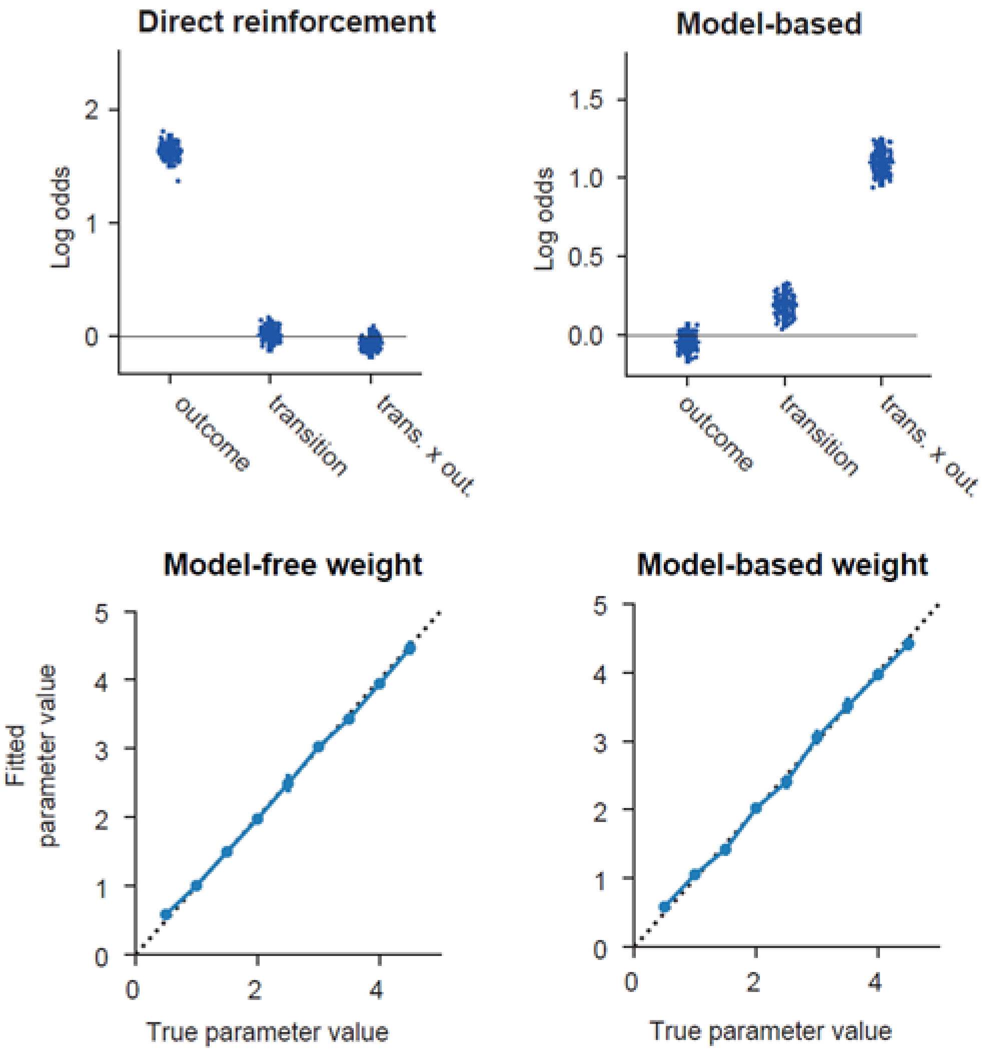
1) The authors have explained their aim to simplify the task, and I understand the importance of having 'good' and 'bad' options and the need to reduce correlations across trials. To me, it appears that the task can be perfectly solved by an MF agent that completely ignores the 2nd state and only tracks the rewards following their actions. Such an agent corresponds to the MF algorithm the authors have used, in the case where the eligibility parameter lambda is set to 1. Indeed, fitted values for lambda are close to 1 (as the authors also acknowledge), which would speak to that. My understanding is that this coupling breaks in the neutral condition - thus MB contributions should be more readily revealed there. My suggestion would be to present the data from the neutral condition separately - I think it has the potential to change the interpretation of the data: if there is already hints of MB contributions to choice in the neutral condition (consistent with what implicitly seems to drive RT already in session 1), then providing subjects with explicit instructions would be amplifying the use of - an already existing - MB strategy. If, on the contrary, even in the neutral condition, no MB contribution was evident, then I would describe the effect of instruction more as breaking an acquired S-R habit.

The reviewer suggests that the task used in our study can be '*perfectly solved*' by a type of model-free strategy usually termed 'direct reinforcement', in which rewards at the end of the trial reinforce the first-step choice taken at the start, irrespective of the second-step state reached. A direct reinforcement strategy (with λ=1) can 'solve' the task in the sense of approximately tracking which first step action has higher reward probability. However, it **will not** produce the same granular pattern of choices as a model-based strategy, and critically **these two strategies are well differentiated** by the analyses we employ for this purpose – the logistic regression analysis of stay probabilities and RL model fitting.

The reason why model-based and direct reinforcement strategies generate different patterns of choices is identical in both the original (Daw et al 2011) two-step task and the simplified version used in the current study. The strategies are differentiated with respect to how the trial outcome (rewarded or not), and state transition following first-step choice (common or rare), jointly influence future choices. For a direct reinforcement strategy, reward at the end of the trial reinforces the first-step choice, irrespective of whether the subsequent state transition is common or rare. This therefore produces a main effect of trial outcome on stay probability but no effect of transition or interaction between transition and outcome. By contrast, for a model-based strategy, reward increases the value of the second-step state where it was obtained, which in turn modifies the value of the first-step actions in proportion to their probability of transitioning to that state. Reward following a rare transition increases the value of the state commonly reached from the not-chosen first-step action, and hence promotes switching on the next trial – the opposite effect to that under a direct reinforcement strategy. A model-based strategy therefore causes the interaction of state transition and outcome to influence stay probability.

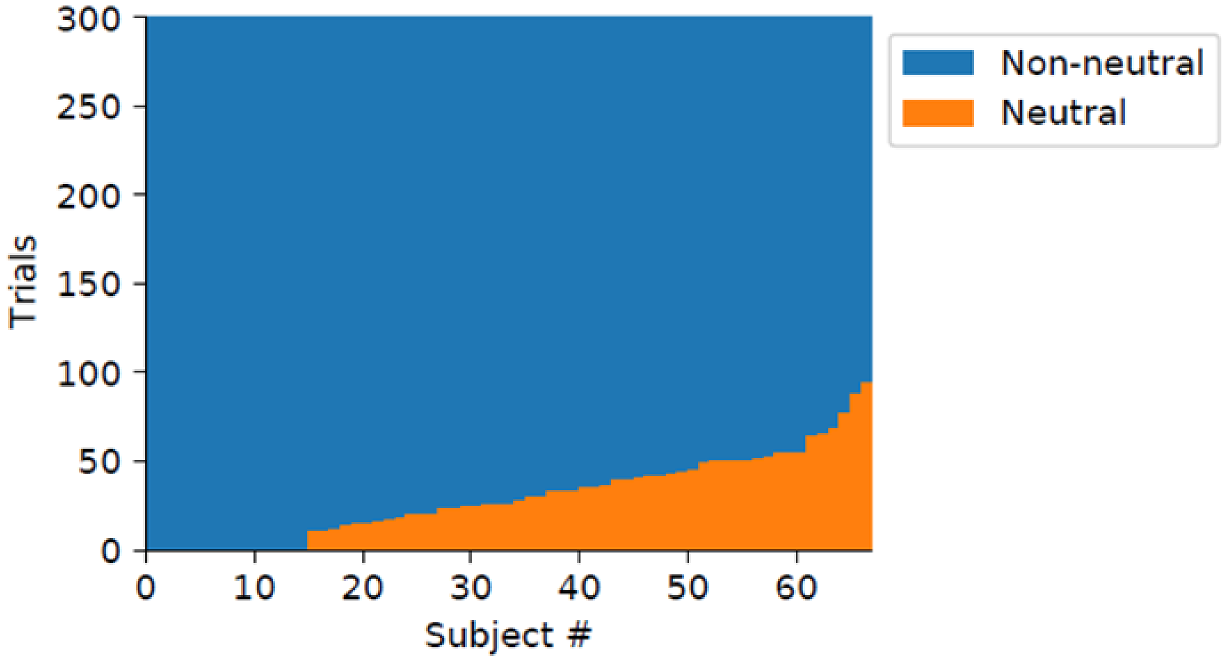
We have previously used simulations to characterise in detail how logistic regression analysis of stay probability can be used to differentiate behavioural strategies in simplified two-step tasks very similar to that used in the current study (Akam et al. PLOS CB 2015). To further verify that this analysis can differentiate these strategies in the current study, we simulated data from both strategies using the exact task used, and analysed the data exactly as in the manuscript (see figure below). As expected, the direct reinforcement strategy generated a strong influence of trial outcome on stay probability and no influence of the transition-outcome interaction, while the model-based strategy produced the opposite pattern.

The second analysis used in the manuscript to differentiate between these strategies was reinforcement learning model fitting. Our parameter recovery analysis (Figure S8) demonstrates that we can accurately quantify the influence of both model-based and model-free strategies on choices in data simulated from the task used in the study. To further confirm that this is the case when the model-free component is direct reinforcement of the first-step choices by trial outcomes, we repeated the parameter recovery analysis for the model-based and model-free weights with the eligibility parameter (λ) set to 1 (see above figure).



Logistic regression and RL model fitting accurately discriminate strategies in simulated data. Top) Logistic regression analysis of stay probabilities for data simulated on the Fixed version of the task using either a direct reinforcement model-free strategy (left panel) or a model-based strategy (right panel). As expected from prior literature (Daw et al. Neuron 2011, Akam et al. PLOS CB 2015), a direct reinforcement strategy produced strong loading on the outcome predictor and no loading on the transition-outcome interaction predictor, while a model-based strategy produced the opposite pattern. **Bottom)** RL model parameter recovery test quantifying how accurately the model-free (left) and model-based (right) weight parameters (i.e. the strength with which each strategy influenced choices) can be recovered from simulated data using model-fitting when the eligibility trace parameter (λ) is set to 1.

Together these data demonstrate that we can effectively differentiate these two strategies using the analyses employed in the paper, when applied to all trials. We therefore do not think that there is a motivation *in principle* to analyse the neutral blocks separately. Additionally, there are strong practical reasons why we think it is unlikely to give insight. Neutral blocks comprise only about 10% of the total trials, and these are unevenly distributed across subjects (see histogram below); due to both the different number of blocks completed and the stochasticity of the block transitions. We therefore do not think it would be appropriate to analyse the neutral blocks independently.



Number of neutral and non-neutral block trials across subjects in session 1 of the fixed task

2) Thank you for including the parameter recovery. I have a question on how the recovery was done. It is stated that 'the parameter under investigation is fixed at the specified true parameter value but the other parameters are drawn randomly for each subject from the population level distributions'. Why are data not generated with the full combination of fitted parameters (e.g. all parameters fixed at their fitted value per participant and all parameters recovered from those artificial data)? In my understanding, this approach would allow an interpretation of the individual model parameters.

Our aim with the parameter recovery analysis was to establish as robustly as possible how accurately we could recover the population mean value of each parameter, as these were the primary measure used for drawing conclusions about learning and debriefing effects within groups, and differences between groups.

Our RL model was hierarchical, with individual subject parameters assumed to be drawn from distributions at the population level. The model-fitting procedure estimated the mean and standard deviation of these population-level distributions for each parameter, and the maximum a posteriori (MAP) parameter estimates for each subject conditioned on these. In our parameter recovery analysis, we felt it was important to evaluate the accuracy of parameter recovery over the plausible range of values each parameter might take, not just at the specific values fit to subjects' behaviour. Therefore, for each parameter we performed simulations where we systematically varied that parameter's value, while drawing the other parameters from the fitted distributions for a given dataset. We drew the other parameters from the fitted distributions rather than using the MAP values for each subject because we needed to repeat each simulation multiple times to estimate the uncertainty in the recovered value, and we were convinced that using values that were not identical on each run, but still matched the distribution of subject's fitted values, was going to give a more conservative estimate. In total the parameter recovery analysis used 1890 separate simulated datasets; 10 repeats for each of 9 systematically varied values for each of 7 parameters, with the remaining parameters for each simulation drawn from distributions fitted to 3 separate experimental datasets. We think this represents a principled and rigorous approach to evaluating the accuracy of parameter recovery.

Decision Letter, third revision:

Our ref: NATHUMBEHAV-200912504C

5th January 2022

Dear Dr. Oliveira-Maia,

Thank you for submitting your revised manuscript "Explicit knowledge of task structure is a primary determinant of human model-based action" (NATHUMBEHAV-200912504C). It has now been seen by the referee who previously had some remaining concerns and their comments are below. As you can see, the reviewer raises no further issues. We will therefore be happy in principle to publish your manuscript in Nature Human Behaviour, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

Marike

Marike Schiffer, PhD

Senior Editor

Nature Human Behaviour

Reviewer #3 (Remarks to the Author):

I thank the authors for patiently explaining the two remaining points that I had admittedly not fully understood. Congratulations on this detailed and thoughtful work.

Author Rebuttal, third revision:

REVIEWER COMMENTS:

Reviewer #3:

Remarks to the Author:

I thank the authors for patiently explaining the two remaining points that I had admittedly not fully understood. Congratulations on this detailed and thoughtful work.

We thank the reviewer for the positive assessment of our response and recognition for the work done.

Final Decision Letter: