Manuscript review (Round 2)

Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context: post-hoc analyses of the components of behavioral flexbility

Dear recommender and authors,

I apologise for the time it has taken me to provide feedback on the paper. It has been a long time since the original submission, so I admit to reading it almost as a new paper. Overall, I noticed an important effort to improve the contextualisation of the study. However, I still find the paper very difficult to follow. This is not only because of the density of the analyses (there are indeed many, and I would encourage the authors to focus on fewer questions for further work), but also because it seems that the paper is a patchwork of somewhat disconnected analyses (first a modelling test, then a test of the two drivers of reversal ability, and finally a test of how behavioural flexibility is related to performance). In addition, I still believe that the paper in its current form cannot be considered a "stand alone" paper. This is due, in my opinion, to (a) many references, without brief summaries, to other work by the team (including previous work from which this paper has been extracted). I understand that one cannot repeat everything said in previous papers, and that this part of the paper was not originally intended to stand alone, but a minimum of additional explanation is needed (e.g. L152 or L490-493). (b) Some parts relevant only to this paper are still unclear for two reasons. First, the explanations are distilled bit by bit when the authors had to use them. This seems efficient for people who know the whole setting, but not for people who are mostly naive to the whole framework. For example, it would be essential to distinguish clearly and sequentially between (1) the empirical setting, (2) the agent-based model, and (3) the statistical models, rather than following the research question titles and adding each of these parts for each question. Second, the authors often assume that the readers are experts in the field or even in their study case/framework. They regularly refer to certain elements without having defined them properly beforehand (e.g. “the parameters” L145, which we assume are the ones described at length without any certainty, “the models”, L165, “the performance”, L147). This is true, in my opinion, for the “background” in the introduction or for the methods (e.g. L152, why is 17 out of 20 for a difference from what is expected by chance?), which makes it difficult to follow the reasoning or even to be critical of it. (c) The paper mixes both “theoretical” and “empirical” efforts. However, a theoretical and an empirical narrative are often different in order to be clearly understood. At present, I feel that the presentation of the theoretical part (e.g. modelling and formulas) lacks a certain formalism. I personally found it very difficult to understand how the modelling works. As a note, I still have mixed feelings about this modelling part overall (see comments on the Bayesian model). I really think that the authors should decide whether this paper should be considered a major “methodological advance” (which I don’t think it is, although it is true that the model may have been tuned for these specific analyses. However, this kind of hand-tuning is often done to adapt to the study system and research questions), or an “empirical test” (which I think it is).

Below, I offer point-by-point comments which I hope will help the authors to see why I have had difficulty in understanding, and therefore evaluating, this piece of research. I admit that I am not the most expert in this area of research, but I hope that these comments will highlight what a more naive reader might find problematic in reading it.

Best regards.

1 MAJOR COMMENTS

Abstract: The emphasis in the presentation is on the methodology used (both the “Bayesian” approach and the combination of simulation and experiments). However, in the presentation of the results, the Bayesian model, the simulation or the experiments used are not explained per se (and sometimes come out of the blue, as in L42-43), nor how they were combined. I would encourage the authors to present their research questions, methods and results in a more step-by-step manner. Moreover, some crucial information is missing (e.g. sample sizes; quantification of the magnitude of effects, such as how the performance of grackles, which is not defined, varies in the puzzle box test). Some terms are also vague (e.g. L35, although somehow understandable after reading the whole paper, the use of “phase” may not be immediately obvious to a novice).

1.1 Introduction

- The first paragraph mixes an explanation of the general context (L51 to 55) with a presentation of methodological approaches (serial reversal experiments). I find this confusing. In this line I would move L49-51 to L55, after the citation. Also, I do not see the point of introducing the experimental approach now, and would move this to paragraph 4, where this methodological background is presented.

- I am confused. The wording in the text suggests that the updating of associations is related to the change and reliability of the environment. A contrario, in the figure it is only related to the reliability of the environment, as if a process (updating associations or sensitivity of associations) is related to an aspect of the environment (change over time and reliability). I would encourage the authors to clarify this. In addition, the authors treat ϕ and λ as separate and even opposite. The problem with this correlation has been raised before, but it is even more glaring here: the current presentation of this background is inconsistent with L494, where this anticorrelation is only "expected", and challenges research question 3), for which ϕ and λ may explain the birds' performance differently (which should be impossible to disentangle if they are indeed negatively correlated). However, as shown in Figure 4, there is a relationship (hence an interaction) of the two parameters together on the performance of the birds. I am sure I am misunderstanding some of this and would encourage the authors to make the link between the compatibility of the statements more explicit.

- A large part of the introduction is devoted to methodology. However, as far as I understand, this is not a methodological paper. I would therefore encourage the authors to increase the description of the biological scenarios of interest (as presented in the second paragraph) and to reduce the methodological description. In addition, it is sometimes unclear why some aspects are discussed at length (e.g. the three "types" of information L116 to 126 - which I would not actually consider as types), as they do not seem to be "exploited" afterwards.

1.2 Research questions

Overall, the research questions are long and the authors could aim for a more concise presentation.
L184 - It is unclear what "abstract" means: how did the models differ from the experiments? Also, I do not understand the argument about differences in estimates between models and experiments. If the models were intended to evaluate the optimal strategy, then it is not necessarily the model structure that may cause differences, but that animals may not behave optimally in nature. Could the authors clarify this statement?
L199 - Why would ϕ and λ not interact to describe the performance of individuals? (as suggested by the "or") Question 3) The prediction is not consistent with the research question (which aims to understand which of the factors explains more variation in performance).
L221 - It seems to me that the authors alternate between behaviours and strategies. I think this is not the same thing (a strategy could consist of one to many behaviours to start with) and the authors should stick to one word.

1.3 Methods

Bayesian model
I find it disconcerting to be presented with the statistical model before the "observational data" (whether from simulation or experiment) are presented. I suggest that the authors restructure the methods in a more classical way, distinguishing between what leads to "data" (simulation scenarios, experimental protocols, etc.) and what is "analysis".

Furthermore, in this model I notice a lack of
- The definition of parameters. For example, the interval for defining association values is not mentioned.
- L305 to 313 As I understand it, modelling association changes across choice events is inevitable of the mechanistic model used (rather than the explanation offered in L305-306). It is therefore not helpful to me (and perhaps to other readers) to have these after the "Implementation" section. The authors might therefore consider structuring the statistical modelling by distinguishing between (1) the background, with the equations, (2) the parameterisation, and finally (3) the implementation (i.e. using Stan/R), and (4) the validation.
- I don't understand what a categorical likelihood means. My understanding of this model, and of mechanistic models in general, is that the equations allow the authors to estimate, for each event and individual, the probability of its realisation, given some values of ϕ and λ. The likelihood can therefore be the product of these probabilities: the best-fitting model (i.e. a particular combination of ϕ and λ) is identified when this product is maximised.
- It is not specified that these models were fitted at the individual level (which they certainly were, as the authors obtained ϕ and λ for each bird). I think this should be specified.
- L281: Is the logit scale for the values obtained positive? From equation 1, it seems to me that the authors use the inverse logit (or so I assume from "logit scale") to constrain a value of ϕ between 0 and 1 (so positive, yes, but not only).

Question 1)
- In these simulations, I don't see the point of having 32 populations if these populations are (almost) perfectly identical, since they can only differ in the range of the parameters being studied? Also, if any combination of ϕ and λ is possible, this violates the expectation/description that both parameters are opposite. Could the authors specify what insights can be gained from modelling unrealistic conditions (e.g. high ϕ and high λ, etc.)? Overall, the explanation of the simulation is not clear enough and relies too much on the previous publication by Logan et al, 2023. I think it is important that the paper can be read without necessarily reading all of the authors' previous work, and would therefore encourage them to provide more detail on this modelling task.
- L346-347: Why should a slope of 1 be assumed? In my opinion, it will give evidence of a correlation, but not necessarily a match, if a is not zero.

Question 3)
In order not to break the flow of the paper, I would encourage authors to include the information on research permissions in a different section (e.g. at the end of the manuscript).

Question 4)
- Authors can continue to refer to birds as "bird" or "j" in the equations.

- The equations seem inaccessible to most readers. For example, the meaning of MVNormal(), L428, LKJcorr() L430, and the associated explanation (I think for the first one) L439-440 is unclear to me.

Question 6)
- What is the unit of latency?
- A correlation between ϕ and λ may (or may not) affect the quality of the fit (i.e. the value of the estimates). This has been raised before, but I may have missed how the authors investigated it (and thus whether it was consequential or not). Also, there are problems with the colours of the curve and the confidence intervals (they are different). Finally, I do not see the point of clustering the trials, even for "clarity".

1.4 Results

- Overall, the readability of the results would be improved by a more concise and 'recognisable' notation (e.g. presentation of confidence intervals between square brackets).

- In my opinion, this first paragraph questions the need to include testing the model as a research question, and I was not fully convinced by the authors' response. As suggested by the other reviewer, if I am not mistaken, the structure of the simulated data and the model are the same, so we simply expect the model to accurately capture the simulation parameters.

- Figure 3: I have a naive question: since the modelling is based on a log-linear relationship, how come the predicted curves can have a sigmoid-like pattern? Is it because of the logit transformation? But I think this curve corresponds to λ (for which there is no such transformation, am I right?).

- Figure 2: I would include the results for both ϕ and λ. Also, I am unclear as to why only 30 individuals are shown when, if I followed correctly, many more were simulated?

- L549-553: These sentences are unclear to me. In particular, I do not understand why this is related to the observed shift, nor how this explanation is actually valid.

- L597-624: What research question is this really referring to? It is supposed to answer Q2, but this was not explained in the associated methods. Also, please be more specific in the manuscript in general. What is "this likely trade-off" here? I understand it to be the one between ϕ and λ, but it seems less clear why this should be likely (despite, I acknowledge, the many previous explanations).

- L689-692: Have the authors explored from simulations the ϕ and λ estimates when considering multiple reversals? Are these also correct?

- L721: I don't understand what this estimate refers to. Is it the one associated with the variable reversal? In this case, it would be good to add the variable being examined for each estimate.

- L711-732: Overall, I am confused, because as I understand it, the first lines (L711 to 722) indicate that there is a homogenisation of ϕ and λ over the experiment between grackles. But then it seems that individuals have different "strategies" overall? Does the strategy refer to the final values or to the changes in ϕ and λ? If so, why is this a "strategy"?

- Figure 7: Is the difference between the curves really significant? I do not know what equation was used to calculate this probability (see my comment in the Minor Comments for the need to give this information), but certainly a confidence interval can be calculated and should be added to the figure.

1.5 Discussion

- L779-807: This is a very long summary of the results. Although appreciated (given the amount of analysis), it would be even more efficient to try to link it to some "ecological" or "behavioural" statement (e.g. avoid mentioning ϕ and λ, but what they represent behaviourally).
- L808-836: It is unclear how these lines discuss results, and it seems more like a justification for such a modelling approach (which would fit in either the introduction or methods). I would encourage the author to reduce this part, perhaps by focusing only on the key technical aspects that have been shown to be necessary for these models to be efficient (i.e. the inclusion of at least one reversal event).
- L845: It is unclear to me for what exactly λ "seems more important" (for performance vs. ϕ?). And what does setting a threshold mean? That λ has an on and off effect, with grackle performance either very low or very high? I do not recall the results being consistent with this view. Can the authors be more specific?

2 MINOR COMMENTS

In my opinion, all equations should be numbered (including those of the statistical modelling, L341-345 and so on) and referenced in the results section when quoting the estimates (otherwise I personally had trouble identifying which equation the authors used). In these equations, I would also stick to either mathematical writing or "reader-friendly" writing (with explicit names of variables, etc.), but not both. Also, what is dnorm(), which is used from equation L501. Is it a normal/Gaussian distribution?
Figure 1: sometimes explore should be always explore, as opposed to "always exploit", shouldn't it?
L69 and 76 - I do not think it is necessary to specify what the letter is in Latin. I would just stick to the Greek version throughout the manuscript.

L100 and 104 - I think comments like “the label of the rate is different from author to author” are more confusing than helpful. I would delete them.
L102 - Please include the numbers of the equations.
L363 - A Poisson distribution assumes that the mean and variance are equal. So have you controlled for the problem of overdispersion? If so, you can mention it. If not, you may want to consider alternative models (e.g. negative binomial; although the modelling here seems robust to the distribution you used, as the results did not change from a Gaussian model).
L443 - This equation is also based on a discrete output variable, so a Gaussian model does not appear to be the best choice. If there is always a decrease in the number of trials (from last to first) then it can be related to count data and could be modelled as a Poisson distribution, I think.
L677 - “... no clear association” between? I assumed number of trials and ϕ and λ, from the brackets, but I would appreciate it if the authors could always be specific. Figure 6: Shouldn’t there be 19 grackle points? I count less (perhaps an imposed limit in the script?) It is also unclear to me how we can deduce the statement of L675-676 from this figure. Could the authors highlight the grackles that also tested for reversal?
L701-702: Could the authors be more specific? I could help a lot by always specifying the magnitude and direction of the effect (here, for example, not ‘changes’ but the ‘increase in’).
L746-748: This should be in the methods, not the results, and linked to the appropriate reference (Logan et al., 2023?).
L779-780: More mechanistic than? And in what sense? And what are the additional findings? This is also mentioned later, but to start the paragraph in this way is, in my opinion, vague and not very informative.

I certainly did not catch all of them, but there were a significant number of typos. Here are a few:

L29 - Species name should be in italics
L61 - Bond et al, 2007 citation format shows problem (square brackets do not match)
L66 - Problem citing Chow et al., 2015
L211 - Problem with citation
L241 - Mathematical notation “belongs to” before the ensemble is missing
L267 to 272 - Double paragraph
L444 - Part related to “a” should go to next line. The same problem exists in L464. L538 - A space is missing after “shifted”.
L594 - Some words are missing/were not deleted? “and the to reach...”
L675 - A period is missing after “(Figure 6)”.
L721 - A space is missing after “+0.17”.
L753 - L762 Correct “$lambda”.
L800 - There are two full stops at the end of the sentence.
L803 - Correct the phi of the Greek letter.
L814 - The quote from Warren is doubled.
L858 - Remove the ‘I’ from ‘IGrackles’.
L876 - Isn’t “grackles” missing after “what strategies”?