REVIEW REPORT ON DOCTORAL DISSERTATION "NEURAL CORRELATES OF PREDICTION ERRORS DURING REWARD AND PUNISHMENT LEARNING" BY KAMIL BONNA

REVIEWER:

Thorsten Kahnt, PhD
Associate Professor
Northwestern University
Chicago, IL, USA

SUMMARY OF THE WORK

The goal of this thesis project is to characterize neural correlates of positive and negative prediction errors in reward-seeking and punishment-avoidance contexts using functional magnetic resonance imaging (fMRI). The first chapters cover the theoretical background of reinforcement learning, magnetic resonance imaging, and network neuroscience. This is followed by a chapter reporting the methods and results of experimental work that characterizes the effects of positive and negative prediction errors on behavior, brain activity, and whole brain network connectivity. This was done using a probabilistic reversal learning task that was framed in either a reward-seeking or a punishment-avoidance context. The behavioral results showed that the learning rate differed significantly between negative and positive prediction errors, but not between reward-seeking and a punishment-avoidance context. Analysis of the fMRI data revealed brain areas in which prediction errors were positively and negatively associated with prediction errors. Moreover, network analyses revealed that integration among large-scale networks was increased during negative compared to positive prediction errors. The findings and limitations of the work are discussed.

STRENGTHS

- The summary of the theoretical background relevant to this work in chapter 1 is excellent. It includes a clear and concise description of classical learning theory and an introduction to reinforcement learning models.
- The commonalities and differences of the Rescorla–Wagner and Temporal Difference learning model are well highlighted.
- The relevant literature on the neurophysiological mechanisms of reward learning, including the dopamine prediction error hypothesis is well summarized.
- The discussion of theoretical problems with punishment-avoidance learning is succinct and adequate.
- The description of the relevant background on the basics of MRI and fMRI in chapter 2 is well compiled and includes an adequate discussion of its shortcomings and limitations.

- The review of the previous literature on fMRI-based investigations into neural correlates of prediction errors is excellent and displays a thorough understanding of the relevant literature.
- One of the key innovative aspects of this work is the emphasis on prediction error-related brain connectivity. Different connectivity methods and previous findings using fMRI connectivity analysis to study prediction error signaling are well presented.
- The introduction to network neuroscience in chapter 3 is well conceived. Specifically, the introduction to graph theory is excellent and covers the key terms and measures to characterize networks. This and the review of the literature on large-scale brain networks during rest and task execution demonstrates a thorough understanding of the relevant literature.
- Chapter 4 covers the experimental part of the thesis: a neuroimaging experiment using a probabilistic reversal learning task. The hypotheses are clearly stated and the experimental design and fMRI data analysis is described in sufficient detail.
- The behavioral probabilistic reversal learning task is well-design and allows for an investigation of neural correlates of positive and negative prediction errors in both reward-seeking and punishment-avoidance contexts. While previous work has addressed similar questions, this provides a sufficiently novel examination into the neural correlates of prediction error signaling.
- A very strong part of the work is the behavioral modeling approach. The comparison of four different models is well justified and implemented, including the methods for model selection and parameter recovery. A particular strength is the Bayesian hierarchical latent-mixture (HLM) parameter estimation procedure.
- The fMRI data are analyzed to identify brain areas in which activity correlates positively or negatively with model-derived prediction errors, and to compare these correlates between reward-seeking and punishment-avoidance contexts.
- The main findings that separate brain systems correlate positively and negatively with prediction errors, and that these correlates are similar in rewarding and punishing contexts provides an important advance to the literature.

- The network analysis constitutes a substantial part of the empirical work. It is also the most novel contribution.
- A methodological strength is that the implementation of beta series connectivity analysis deviates from its original form by estimating trial-specific responses using trial-specific GLMs, as proposed by Mumford et al. (Mumford et al., 2012).
- The key finding from the network analysis is that the configuration of the whole brain network differs between positive and negative prediction errors but not so much between reward-seeking and punishment-avoiding contexts.
- The discussion section provides a mostly appropriate interpretation of the findings in the light of dual system, reference point, and Global Workspace hypotheses.

WEAKNESSES

- There appears to be a disconnect between the way positive and negative prediction errors are conceptualized in the behavioral and connectivity analyses on the one hand, and model-based fMRI analyses on the other. Whereas the behavioral and connectivity analyses compare trials with positive and negative prediction errors, the model-based fMRI analysis does not consider the sign of prediction errors but tests for positive and negative correlations with model-derived prediction errors, spanning a continuum from negative to positive prediction errors. This is obviously not the same as asking what brain areas encode positive and negative prediction errors. The analysis as implemented cannot reveal brain areas that encode positive and/or negative prediction errors. This could only be done by separately testing for prediction error-related parametric modulations within trials with (1) positive and (2) negative prediction errors. Such an analysis could have four possible outcomes: (1) regions correlating positively with positive prediction errors, (2) regions correlating negatively with positive prediction errors, (3) regions correlating positively with negative prediction errors, (4) regions correlating negatively with negative prediction errors.

- Because prediction errors in the model-based fMRI analysis are not separated by sign but span a continuum from negative to positive, this analysis cannot test whether positive and negative PEs are encoded in different networks. Insofar, the statement on page 101 that "*I observed a clear distinction between dopaminergic, striatal system signaling positive prediction errors and insular-frontal system signaling negative prediction errors*" is in no way supported by the data.
- Related to the two points above, the model-based fMRI analysis confounds prediction error-related signals with the valence of the outcome. Prediction errors for positive outcomes (gains in the reward-seeking context and neutral outcomes in the punishment-avoidance context) are always positive, whereas prediction errors for negative outcomes (neutral outcomes in the reward-seeking context and losses in the punishment-avoidance context) are always negative. Thus, positive and negative correlations with PEs shown in Figure 4.5. are confounded by signals that may have nothing to do with continuous prediction errors but categorically differ between positive and negative outcomes. This confound complicates the interpretation of the findings presented in Figure 4.5. To avoid this confound, the prediction error regressor in the 1st level GLM has to be orthogonalized with respect to a dummy regressor that codes for outcome valence (i.e., good vs. bad outcome).
- Correlations between activity and model-derived prediction errors may only capture part of the key features of prediction errors, and thus may lead to the erroneous conclusion that an area codes prediction errors. This fallacy is avoided in the axiomatic testing approach proposed by Rutledge and Glimcher (Rutledge et al. 2010). This caveat of a correlative approach deserves to be discussed.
- The reinforcement learning model incorporates value updating for both the chosen and unchosen option (equation 4.2). This explicitly incorporates knowledge about the anti-correlated task structure into the model. It is well-established that such models fit data better than standard RL (Hampton et al., 2006), but it represents a deviation from standard RW and TD learning and this deserves to be discussed.

EVALUATIVE SUMMARY

Overall, this is a very strong doctoral dissertation. It manages to provide a substantial advance in a crowded area of research. A particular strength of the written dissertation are the description of the theoretical background, which demonstrates fundamental theoretical knowledge in the relevant areas. The behavioral modeling and the network analyses demonstrate mastery and innovation in the independent application of this theoretical knowledge to new scientific problems. This results in interesting findings that will move the field forward. There are weaknesses in the approach of the model-based fMRI data analysis that affect the interpretation and conclusions. However, these weaknesses are relatively minor compared to the impressive strengths of the other parts of the work. Overall, it is my assessment that the dissertation meets the conditions set out in Art 187 paragraph 1 and 2 of the Act on Higher Education and Science of July 20, 2018 (as amended).

APPLICATION FOR DISTINCTION

Based on the overall quality or the written dissertation, the behavioral modeling, and the novel investigation of how prediction error signaling modulates whole-brain network connectivity, I believe the doctoral dissertation deserves a distinction.

REFERENCES

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci *26*, 8360-8367.

Mumford, J.A., Turner, B.O., Ashby, F.G., and Poldrack, R.A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage *59*, 2636-2643.

Thorsten Kahnt, PhD 3/25/2022