Prof. dr hab. Daniel Krzysztof Wójcik
Laboratory of Neuroinformatics
Nencki Institute of Experimental Biology
of the Polish Academy of Sciences
02-093 Warsaw, Pasteur St. 3, Poland

Warsaw, April 4th, 2022

Review of the doctoral thesis by Mr. Kamil Bonna entitled
"Neural correlates of prediction errors during reward and punishment learning"

Learning from trial and error is a common paradigm used by animals to cope with environment but also a useful approach in machine learning. Despite years of study specific abstract models used by the brain and the implementation details in different context are not yet fully understood. In his thesis, Mr Kamil Bonna considers probabilistic reversal learning task using behavioral modeling and functional magnetic resonance studies, where he looks both at linear model descriptions and network analysis.

This is an extensive work of around 150 pages including six chapters, bibliography, and an appendix with supplementary information. The text is adequately illustrated with over 20 figures and 6 tables. Extensive bibliography contains close to 300 items. The thesis is prepared neatly, the figures are generally readable, and the editorial quality is very high, I found only a few minor typos, no major errors.

The thesis starts with a brief introduction where the author announces the main challenges he wants to tackle, that is answering the dilemma of a common system or two independent systems responsible for handling positive and negative prediction errors, the difference between reward and punishment learning, and the reference hypothesis. The main strategy of attack on three different levels is presented and the structure of the thesis is announced.

This is followed by three extensive chapters, introducing the basic concepts of reinforcement learning, human brain imaging of prediction errors, and network neuroscience. In the first chapter the author briefly reviews reinforcement learning as a general psychological problem, presents models used to describe it, experimental evidence for its neural implementation as well as competing hypotheses addressing reward and punishment learning. This forms the conceptual basis for the thesis. Chapter 2 introduces the basics of MRI physics and fMRI studies of prediction-error related activity and functional connectivity. In Chapter 3 the author presents the elements of the network theory and the studies of functional brain networks in resting state and during cognition. The overview of the relevant research literature is very extensive and definitely more than adequate to satisfy the demands of a PhD thesis.

Extensive chapter 4 presents the results of the different approaches followed by the author. This starts with another introduction followed by presentation of a range of hypotheses which the author has tested. These are 1. processing of positive and negative prediction errors is handled by two separate systems; 2. the valence of outcome is relative to the context; 3. more integrated network is needed to process negative rather than positive prediction errors. These main hypotheses are then discussed and expanded providing a set of questions the author had addressed in his work.

Once the hypotheses are revealed the author presents his experimental paradigm including data preprocessing. He then goes into details of his behavioral modeling, including

Bayesian models, Markov Chain Monte Carlo estimation, before going into behavioral performance and model selection.

The second approach used is model-based fMRI analysis using Generalized Linear Models and the analysis of context-dependent and context-independent processing of prediction errors.

The final approach of which the author is particularly proud is the analysis of functional brain networks. The author discusses here the technicalities of the methods, such as possible dependence of results on different brain parcellations and his choices and adaptation to standardize the analysis across subjects, construction of network, the study of its modularity and community structure, and large scale networks interactions, as well as stress-tests or robustness precautions using between-community agreement and consensus networks.

Chapter 4 is closed by a discussion, conclusions, limitations, and followed by a summary in a separate chapter.

Overall I do not have major complaints, the description of the experimental and analytical procedures were to me adequate and results well documented. I had several questions while reading but practically all of them have been addressed by the author somewhere. One issue that remains regards the availability of code, and to some extent the data. While I think the contribution of the author to the problem of neural implementation of reward and punishment learning in humans is significant, it will probably not end the discussions in the community. I think it would be valuable to repeat the author's complete protocol, including behavioral modeling with studies of neural activity and functional connectivity for other tasks or for other modalities, for instance using EEG, MEG, fNIRS or intracranial studies on epileptic patients, instead of fMRI. While the description of the author is sufficient for reproduction, the amount of work needed to put in place the complete workflow seems quite demanding. It would be useful to have access to the code to avoid duplication of this work. Further, adding the access to the collected data would make it possible for others to reproduce the author's results but also to test the workflow before applying to one's own data. I wonder what are the author's plans and views on this or if anything has been done towards that. I am also curious if the author plans to himself develop the project along any of the directions just mentioned.

Minor comments

The organization of the text within chapter is fine, but overall organization of the thesis I found cumbersome. For reasons unknown, the first and last chapters – introduction and summary – are not numbered. The first three chapters focusing on reinforcement learning, human brain imaging, and network neuroscience, combine motivational parts which would be natural in the introduction or in the discussion with clearly methodological issues. Chapter 4 contains all the results of the author, however, they are also combined with methods used by the author, both experimental and analytical, and followed by a discussion. This is a minor comment, however, the provided content naturally falls into five chapters, which should all be numbered or none: Introduction, Methods, Results, Discussion, Summary. Present chapters 1 to 3 should be largely sections within Methods, in the same way as present chapter 4 is organized. Some historical, overview and motivational parts could be moved to the introduction. Sections 4.7-4.9 should be extracted into a separate chapter Discussion, which would close with the present Summary. The split into discussion, conclusions, limitations and summary, seemed a little overdone to me, especially on the structural level. The present structure made me feel the thesis was imbalanced with too much space devoted to introductory and overview material as compared to own results.

I found it rather inconvenient to find the main hypotheses the author wants to test presented as late as on page 53. It would not hurt to reveal them to the reader a little earlier, say by page 3 of the introduction, in such a coherent and readable form. In a sense, they made it to the introduction but only page 53 made them comprehensible to me.

To summarize, despite my minor criticism, **I regard the doctoral thesis by Mr Kamil Bonna very highly**. I believe it is a significant contribution to the problem of neural implementation of reward and punishment learning in humans. I think his approach, addressing the problem combining behavioral modeling with studies of neural activity and functional activity of the same subject is very original and provides a coherent picture of the neural underpinnings of the studied problem. **In my opinion this thesis satisfies all the usual and formal demands set out for** doctoral theses and I strongly support awarding doctoral degree to Mr Kamil Bonna. [Uważam, że rozprawa doktorska mgr. Kamila Bonny spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668), dlatego zwracam się do Wysokiej Rady Dyscypliny Naki Fizyczne Uniwersytetu Mikołaja Kopernika o dopuszczenie mgr. Kamila Bonny do dalszych etapów przewodu doktorskiego.]

I think the breadth of approach and the completeness of presentation deserve a distinction and I am willing to recommend it to the Discipline Board in Physical Sciences of the Nicolaus Copernicus University. However, before voting in this matter, I would like to know the publication status of these results as I was unable to find them in the literature.

Prof. Daniel K. Wójcik