How your brain learns penguins are birds - neural patterns of exception learning

Rebekka Heinen (rebekka.heinen@ruhr-uni-bochum.de)

Robert Lech (robert.lech@ruhr-uni-bochum.de)

Boris Suchan (boris.suchan@ruhr-uni-bochum.de)

Nikolai Axmacher (nikolai.axmacher@ruhr-uni-bochum.de)

Department of Neuropsychology, Ruhr University Bochum
Germany

Abstract

We know that, despite their bird-like appearance, bats are mammals. How does the brain learn such exceptions? We conducted an fMRI study where participants categorized stimuli, including exceptions. We analyzed learning behavior using prototype and exemplar learning models. Initially, both models equally predicted behavior, but the exemplar model later outperformed the prototype model. Combining stimulus similarity patterns from the learning models with a whole-brain searchlight using representational similarity analysis (RSA), we found that the prototype model matched neural patterns in frontal and temporal regions, while the exemplar model matched patterns in the visual cortex. Our findings support the idea of flexible recruitment of multiple learning systems involved in concept learning, based on task demands.

Introduction

Concept learning involves acquiring new knowledge and forming category representations, such as understanding that birds have wings and can fly. But what happens with exemplars that share fewer features with the category prototype, like penguins (Fig. 1A)? There is an ongoing debate if single system approaches, like prototype learning (categories represented by their central tendency based on shared features) and exemplar learning (categories represented by specific exemplars; Zeithamova et al. (2019)) can be used simultaneously, depending on task demands (Minda, Roark, Kalra, & Cruz, 2024). We can explore this by manipulating participant learning strategies, focusing on concept exceptions. These exceptions should push participants to use multiple learning systems, as prototype learning will lead to false classifications of exceptions. Using representational similarity analysis (RSA; Kriegeskorte, Mur, and Bandettini (2008)) on learning models and in a whole-brain searchlight, we can test how learning models and the human brain represent concepts and their exceptions in feature space. Prior studies compared models to behavioral similarities, including exceptions (Heffernan, Schlichting, & Mack, 2021; Xie & Mack, 2024), or to neural data without exceptions (Heffernan, Adema, & Mack, 2021). We aim to fill this gap by testing how learning models reflect how the human brain learns categories and their exceptions using RSA to compare model and brain similarity patterns.

Methods

The study was approved by the local ethics committee and conducted in accordance with the declaration of Helsinki. We tested 26 participants (M_{age}=24.57, SD_{age}=2.70, 15 female) undergoing functional magnetic resonance imaging (3 Tesla; voxel size = 1.65 x 1.65 x 5 mm; TR = 2.2 s) during a categorization learning task. Across five blocks with 98 trials each, participants sorted a stimulus into one of two categories (Fig. 1B), receiving feedback on their category decision after each trial. Stimuli were adapted from Cook and Smith (2006). Depending on their colors, stimuli would either belong to category A or B. While the two concept prototypes differed in all dimensions, typical exemplars shared most of their colors with the prototype of their class, except for two stimuli that shared most of their features with the other class (exception). Using these stimuli, where participants had to prior knowledge, we could analyze how the brain acquires concept knowledge, but also how it handles concept exceptions. First, we tested for learning performance differences between stimulus types using ANOVA and post-hoc paired t-tests. We then trained a prototype model (Minda & Smith, 2012) and an exemplar model (Nosofsky, 2012) to predict participant responses, indicating if participant choices were based on prototype or exemplar learning. We computed root mean square deviation as a measure of model fit to participants' data and compared model performance using paired t-tests. All results are corrected for multiple comparisons using Bonferroni. We then extracted similarity matrices from the two participant specific models. In the prototype model, we extracted the similarity of each stimulus to the prototype of class A and of class B for each block. For the exemplar model, we extracted the similarity of each stimulus to all other exemplars of class A and B (Fig. 1E). In a whole-brain searchlight, we computed the corresponding similarities from neural data (prototype x stimulus, stimulus x stimulus matrices per block) and correlated the similarity of the model and brain matrices (Spearman-rho) at each searchlight sphere location, resulting in a similarity brain map for each model. Using randomise in FSL, we tested for significant differences between the prototype and the exemplar model similarity brain maps.

Results

Flexible recruitment of learning strategies

While prototypes and typicals are learned early, participants fail to categorize the exceptions correctly in the beginning of the experiment (main effect stimulus type: $F_{(2,50)}$=5.92, p=0.017; main effect block: $F_{(4,100)}$=13.58, $p<0.001$; interaction: $F_{(8,200)}$=4.21, p=0.008; block 2 exceptions vs. prototypes: $t_{(25)}$=-4.20, p=0.004; block 2 exceptions vs. typicals: $t_{(25)}$=-3.43, p=0.030; all others $p>0.05$; Fig. 1C). This distinction between the early and late blocks is also present when comparing learning model performance: While the models do not differ in the first block, starting from block three the exemplar model outperforms the prototype model, suggesting that participants start to employ exemplar learning (block 1: $p>0.05$; 2: $p>0.05$; 3: $t_{(25)}$=3.52, p=0.008; 4: $t_{(25)}$=5.09, $p<0.001$; 5: $t_{(25)}$=5.06, $p<0.001$; Fig. 1D).

Distinct networks reflect learning model similarities

Our whole-brain searchlight revealed distinct networks for the prototype and the exemplar model, reflecting the learning model similarities across all five blocks (Fig 1F). While the exemplar model is linked to neural similarities in visual areas including the lateral occipital cortex and lingual gyrus (cluster 1: $t_{(25)}$=7.39; cluster 2: $t_{(25)}$=6.35), the prototype

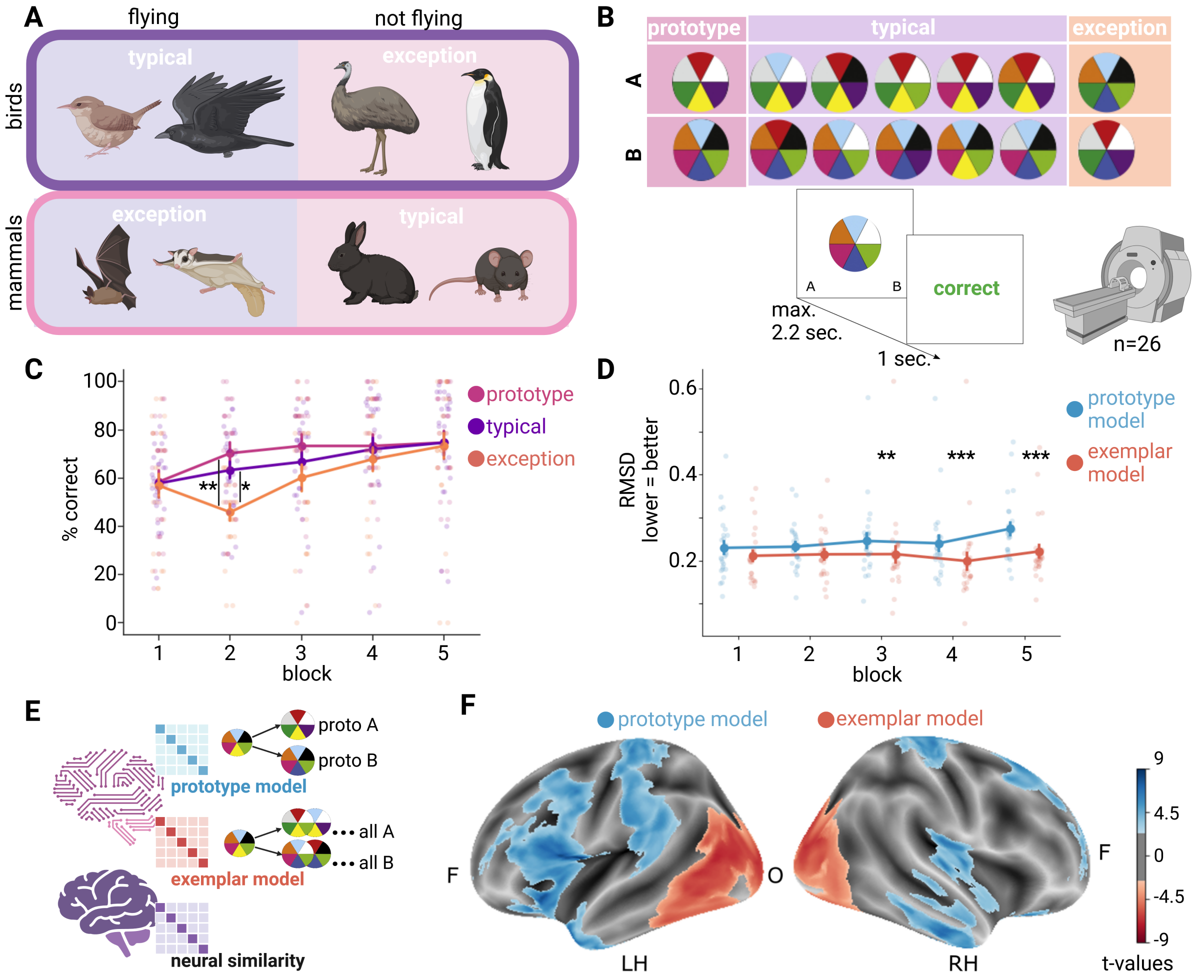

Figure 1: A) Examples for prototypes and exceptions. B) Paradigm. C) Percent of correct participant ratings for prototypes, typicals and exceptions across the five blocks. D) Root mean square deviation (RMSD); smaller values indicate a better model fit to participants learning performance. E) Overview of RSA procedure. F) Searchlight results. Error-bars represent standard errors; Points represent single participants; *p<0.05; **p<0.01;***p<0.001; Created in https://BioRender.com.

model reflects similarities in fronto-temporal regions such as the inferior frontal and middle temporal gyrus and the posterior cingulate (cluster 1:$t_{(25)}$=4.95; cluster 2: $t_{(25)}$=3.81;cluster 3: $t_{(25)}$=3.56). Results are FWE corrected using TFCE with p_{corr}<0.01 threshold.

Outlook

We show flexible recruitment of both prototype and exemplar learning in line with a multiple systems approach, with the two systems being reflected in distinct brain networks in line with current literature (Bowman, Iwashita, & Zeithamova, 2020; Minda et al., 2024). Surprisingly, we did not find hippocampal involvement, a region crucial for concept learning (Mok & Love, 2023; Mack, Love, & Preston, 2016; Bowman & Zeithamova, 2018). Thus, our current analysis focuses on the hippocampus as a region of interest. In addition, we have collected a behavioral follow-up including only prototypes and typicals, to test whether it is indeed the presence of exceptions leading to the better fit of the exemplar model. We will also test an additional learning model which covers both prototype and surprising events (Love, Medin, & Gureckis, 2004). Lastly, in an additional searchlight, we will test whether exceptions build their own subcategory (is a penguin more similar to a kiwi because their are 'odd' birds?), testing if exceptions become more similar to each other compared to their own category exemplars.

Acknowledgements

R.H., R.L., B.S. and N.A. received funding from the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation)—project no. 122679504—SFB 874, the European Research Council (grant: CoG 864164 to N.A.) and the German Mercator Research Center Ruhr (grant: Ex-2021-001 to R.H.).

References

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, *9*. doi: 10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614.

Cook, R. G., & Smith, J. D. (2006). Stages of abstraction and exemplar memorization in pigeon category learning. *Psychological science*, *17*(12), 1059–1067. doi: 10.1111/j.1467-9280.2006.01833.x

Heffernan, E. M., Adema, J. D., & Mack, M. L. (2021). Identifying the neural dynamics of category decisions with computational model-based functional magnetic resonance imaging. *Psychonomic Bulletin & Review*, *28*(5), 1638–1647.

Heffernan, E. M., Schlichting, M. L., & Mack, M. L. (2021). Learning exceptions to the rule in human and model via hippocampal encoding. *Scientific reports*, *11*(1), 21429. doi: 10.1038/s41598-021-00864-9

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4. doi: 10.3389/neuro.06.004.2008

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological review*, *111*(2), 309–332. doi: 10.1037/0033-295X.111.2.309

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208.

Minda, J. P., Roark, C. L., Kalra, P., & Cruz, A. (2024). Single and multiple systems in categorization and category learning. *Nature Reviews Psychology*, *3*(8), 536–551. doi: 10.1038/s44159-024-00336-7

Minda, J. P., & Smith, J. D. (2012). Prototype models of categorization: basic formulation, predictions, and limitations. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 40–64). Cambridge University Press. doi: 10.1017/CBO9780511921322.003

Mok, R. M., & Love, B. C. (2023). A multilevel account of hippocampal function in spatial and concept learning: Bridging models of behavior and neural assemblies. *Science advances*, *9*(29), eade6903. doi: 10.1126/sciadv.ade6903

Nosofsky, R. M. (2012). The generalized context model: an exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 18–39). Cambridge University Press. doi: 10.1017/CBO9780511921322.002

Xie, Y., & Mack, M. L. (2024). Reconciling category exceptions through representational shifts. *Psychonomic Bulletin & Review*, *31*(6), 2621–2633.

Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., Van Kesteren, M. T., & Wutz, A. (2019). Brain mechanisms of concept learning. *Journal of Neuroscience*, *39*(42), 8259–8266.