Learning exceptions to the rule in human and model via hippocampal encoding

Emily M. Heffernan (emily.heffernan@mail.utoronto.ca)
Michael L. Mack (michael.mack@utoronto.ca)
Department of Psychology, University of Toronto
100 St George St, Toronto, ON M5S 3G3 Canada

Abstract

We explore the impact of learning sequence on performance in a rule-plus-exception categorization task. Behavioural results indicate that exception categorization accuracy improves when exceptions are introduced later in learning, after exposure to rule-following stimuli. Simulations of this task using a neural network model of hippocampus and its subfields replicate these behavioural findings. Representational similarity analysis of the model's hidden layers suggests that model representations are also impacted by trial sequence. These results provide novel computational evidence of hippocampus's sensitivity to learning sequence and further support this region's proposed role in category learning.

Keywords: category learning; learning sequence; hippocampus; computational modelling

Introduction

Category learning is a mechanism by which we make sense of the influx of information present in our daily lives. When we encounter a novel object or situation, we can compare it to previous experiences to make inferences about its qualities. Research on category learning has been an active topic of exploration in cognitive sciences for decades, as evidenced by a dense body of literature that explores implicated brain regions, underlying neural processes, the nature of representations used to store categorical information, and so on. Although emerging work continues to shed light on the many facets that underlie this dynamic cognitive skill, many questions remain surrounding the impact of the learning process itself on category learning.

Although category learning is a complex process that recruits multiple brain regions (for a review, see Zeithamova et al., 2019), recent work has implicated hippocampus in this cognitive process (Bowman & Zeithamova, 2018; Davis, Love, & Preston, 2012a, 2012b; Mack, Love, & Preston, 2016; Schapiro, McDevitt, Rogers, Mednick, & Norman, 2018). Notably, hippocampus forms conjunctive representations that bind together multiple features (O'Reilly & Rudy, 2001; Sutherland & Rudy, 1989), and this rapid formation of conjunctive representations is especially important when learning complex category structures. An example of such a problem is a rule-plus-exceptions task in which most stimuli adhere to a rule, but a small subset of exceptions violate this rule. The learner must detect general patterns while also distinguishing and remembering irregularities. Conjunctive representations are crucial to learning exceptions in rule-plus-exception categorization tasks (Davis et al., 2012a; Love & Gureckis, 2007). A wide body of evidence spanning several literatures indicates enhanced memory for these schema-violating exception items (e.g., Goodman, 1980; Palmeri & Nosofsky, 1995; von Restorff, 1933). This memory advantage may be attributed to the formation of more detailed neural representations (Davis et al., 2012a).

The formation of distinct conjunctive representations adheres to traditional views of hippocampus's role in episodic memory (e.g., McClelland, McNaughton, & O'Reilly, 1995), but recent work has implicated hippocampus in rapid statistical learning (Schapiro, Turk-Browne, Botvinick, & Norman, 2017). Hippocampus's ability to support these complementary processes may be attributed to two white matter pathways that traverse specialized hippocampal subfields. Dentate gyrus (DG) and cornu ammonis 3 (CA3) fall along the trisynaptic pathway (TSP) and are associated with sparse, pattern-separated representations; conversely, CA1, which is part of the monosynaptic pathway (MSP), employs dense, overlapping representations ideal for extracting regularity (for a review, see Duncan & Schlichting, 2018). The pattern separation and pattern detection enabled by MSP and TSP render hippocampus well-suited to support the divergent needs of rule-plus-exception learning.

Studies on populations with limited hippocampus function further emphasize this brain region's importance to rule-plus-exception learning. Individuals with underdeveloped or damaged hippocampus exhibit impaired rule-plus-exception learning, likely due to their reduced ability to form the requisite conjunctive representations (Love & Gureckis, 2007). However, limited work has explored how performance in a rule-plus-exception task might be manipulated in healthy young adults. In this work, we aim to explore how learning can be influenced not by the structural characteristics of hippocampus but rather by the order in which information is presented.

Previous work on sequence manipulation has explored how transitions between trials impact category learning. Carvalho and Goldstone's (2015) sequential attention theory (SAT) posits that a blocked design wherein items from one category are presented in succession, followed by a block of members from the opposing category (AAABBB), emphasizes intra-category differences (i.e., differences between members of the same category), and an interleaved design that switches frequently between categories (ABABAB) emphasizes inter-category differences (differences between members of opposing categories). Mathy and Feldman (2009) have explored the impact of delaying the introduction of exception items from one category. They found that this rule-based manipulation significantly improved learning outcomes compared to sequences that maximized or minimized the similarity between stimuli in consecutive trials.

To further explore the impact of learning sequence on categorization, and to test how this manipulation targets hippocampus, we devised a behavioural task in in which the introduction of exception items occurred either early in learning or later in learning after exposure to rule-following items. We predicted that learners who were introduced to exceptions later in learning would develop a better initial understanding of category structure, which in turn would improve their ability to detect and learn to categorize these "late exceptions." To test hippocampus's sensitivity to this behavioural manipulation, we next ran this task on a neural network model of hippocampus (Ketz, Morkonda, & O'Reilly, 2013; Schapiro et al., 2017). We predicted that if hippocampus was sensitive to our manipulation, the model would also be better able to categorize exception items when their introduction was delayed. Finally, we conducted a representational similarity analysis (RSA) of hidden layers of the neural network to explore how this manipulation impacted the model's "neural" representations. Because surprise facilitates pattern separation (Davis et al., 2012b; Yamaguchi, Hale, D'Esposito, & Knight, 2004), we expected to observe more granular, pattern-separated representations of exceptions introduced later in learning.

Behavioural Experiment and Analysis

Methods

Participants All participants were University of Toronto students who received course credit for participating. Data were collected in-lab and, as necessitated by COVID-19-imposed remote study, online. There were 49 in-lab participants (37 females; mean age 19.1 years, SD 3.3 years) and 44 online participants (20 females, 2 other; mean age 19.9, SD 1.0 years). In total, 93 participants completed the experiment. All procedures were conducted in accordance with the University of Toronto's Research Ethics Board.

Stimuli Throughout the experiment, participants viewed 10 images of flowers. Flower stimuli had four binary features: outer petal colour, outer petal shape, inner petal shape, and central disc colour (Figure 1). The central disc was determined to be least salient in a norming study and as such was chosen to be non-diagnostic and varied randomly between stimuli. Outer petal shape was the most salient dimension. Category structure was assigned using the Type 3 problem defined by Shepard, Hovland, and Jenkins (1961). Based on this structure, each stimulus was characterized as one of three types. Two stimuli were defined as category prototypes and were maximally dissimilar across all three diagnostic dimensions. Rule-followers differed from their category's prototype by one diagnostic dimension, and exception stimuli differed from their category's prototype by two dimensions (i.e., they were more similar to the prototypes of the opposite category).

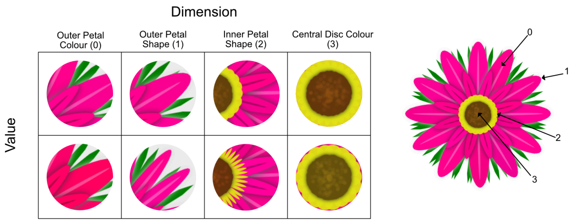

Figure 1. Flower stimuli had four binary features.

Procedure Participants completed three learning blocks, each with 48 trials. Full feedback was provided after each trial. Participants were randomly assigned into one of two conditions. In the early exceptions (EE) condition, participants viewed exceptions in all three learning blocks, and in the late exceptions (LE) condition, participants were not exposed to exceptions until the second learning block (Table 1). Participants saw two times more prototypes than rule-followers and exceptions to anchor each category. Following the learning blocks, participants completed a test block with 48 trials.

Table 1. Stimulus distribution across blocks/conditions. "E," "P," and "R" denote exceptions, rule-followers, and prototypes, respectively.

Block	EE Condition	LE Condition
Learning 1	24P/12R/12E	24P/24R/0E
Learning 2	24P/12R/12E	24P/12R/12E
Learning 3	24P/12R/12E	24P/0R/24E
Test	16P/16R/16E	16P/16R/16E

Results

Participants were excluded from the analysis if they failed to achieve an accuracy of over 0.75 for at least one stimulus type in any of the learning or test blocks or if over 20% of their reaction times fell outside the range [0.15s, 2s]. These criteria were chosen a priori to exclude participants who were not putting effort into learning the category structure. Based on these criteria, 10 SONA participants and two in-lab participants were excluded, resulting in a total of 81 participants who were included in further analyses. Data from included participants were further cleaned to exclude any responses less than 0.15s or greater than 2s (7.1% of all trials were excluded).

For the learning blocks, effects of stimulus type (exception, prototype, or rule-follower), condition (LE or EE), and repetition on accuracy were first assessed. Here, repetition is defined as the number of appearances of a given stimulus type throughout the three learning blocks. To account for the higher number of prototypes versus other stimulus types, only the first 36 repetitions for each type were included in analysis. A general linear mixed-effects (GLME) model was fit to the learning data (using the lme4 package v. 1.1–26 in R v. 4.0.4) to predict trial-by-trial accuracy. Inputs to the model were trial scores (0 for incorrect, 1 for correct). This model included stimulus type, condition, and repetition as fixed effects and participant as a random effect.

In the LE condition, there was a main effect of repetition: categorization accuracy for exceptions, prototypes, and rule-followers improved significantly with repetition (Figure 2; $\beta_E = 0.038$, $P < .001$, 95% CI [0.027, 0.048]; $\beta_P = 0.037$, $P < .001$, 95% CI [0.0239,0.052]; $\beta_R = 0.019$, $P < .001$, 95% CI [0.009, 0.030] – where subscripts e, p, and r denote exceptions, prototypes, and rule-followers, respectively). In the EE condition, categorization accuracy for prototypes improved significantly with repetition ($\beta_P = 0.017$, $P = .011$, 95% CI [0.004, 0.030]); however, categorization accuracy for exceptions and rule-followers showed no significant improvement with repetition ($\beta_E = 0.006$, $P = .275$, 95% CI [-0.005, 0.016]; $\beta_R = -0.010$, $P = .081$, 95% CI [-0.020, 0.001]). There was also an interaction between repetition and condition. Categorization accuracy improved more with increased repetition in the late condition than in the early condition for all stimulus types ($\beta_E = 0.032$, $P = < .001$, 95% CI [0.0173, 0.0469]; $\beta_P = 0.021$, $P = .036$, 95% CI [0.00141, 0.0403]; $\beta_R = 0.029$, $P < .001$, 95% CI [0.0136, 0.0438]).

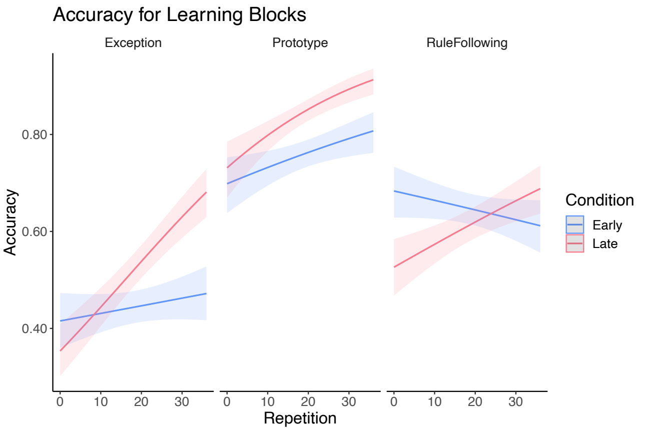

Figure 2. Estimated marginal means from the statistical analysis of categorization performance in learning blocks. Bands indicate standard error.

The test block was also analyzed using a GLME model (Figure 3). This model had the same predictors as the learning block model but without repetition as a fixed effect. In the test block, categorization accuracy was above chance for prototypes in the LE and EE conditions ($\beta0_{P\text{-}LE} = 2.492$, $P < .001$, 95% CI [2.169, 2.835]; $\beta0_{P\text{-}EE} = 2.010$, $P < .001$, 95% CI [1.718, 2.315]), for rule-following items in the LE and EE conditions ($\beta0_{R\text{-}LE} = 0.996$, $P < .001$, 95% CI [0.749, 1.250]; $\beta0_{R\text{-}EE} = 0.914$, $P < .001$, 95% CI [0.666, 1.166]), and for exceptions in the LE condition ($\beta0_{E\text{-}LE} = 0.732$, $P < .001$, 95% CI [0.490, 0.980]). Performance for exceptions in the EE condition was not significantly above chance ($\beta0_{E\text{-}EE} = 0.132$, $P = .272$, 95% CI [-0.107, 0.371]). Categorization accuracy was significantly higher in the LE condition than in the EE condition for prototypes ($\beta_P = 0.482$, $P = .034$, 95% CI [0.040, 0.931]) and exceptions ($\beta_E = 0.600$, $P = .001$, 95% CI [0.260, 0.944]). There was no significant difference between conditions for rule-followers ($\beta_R = 0.082$, $P = 0.645$, 95% CI [-0.269, 0.437]).

Preliminary Discussion

The results from both the learning and test blocks indicate that manipulating trial order by delaying the introduction of exceptions significantly impacted categorization accuracy. As hypothesized, categorization accuracy for exception items improved more over time in the LE condition than in the EE condition. However, the impacts of learning sequence extended beyond exception stimuli. In the LE condition, performance improved with repetition for all stimulus types; conversely, in the EE condition, performance only improved significantly over time for prototypes, and this improvement was still less than improvement in prototype learning in the LE condition. These differences in performance were also apparent after learning, in the test block. Categorization accuracy for prototypes and exceptions was significantly higher for participants who had previously been exposed to the LE sequence. These results indicate that the opportune moment to introduce exceptions in a rule-plus-exception task seems to be after participants have been familiarized with the general category structure. To further explore how our manipulation impacted neural representations, we next simulated this task using a neural network model of hippocampus.

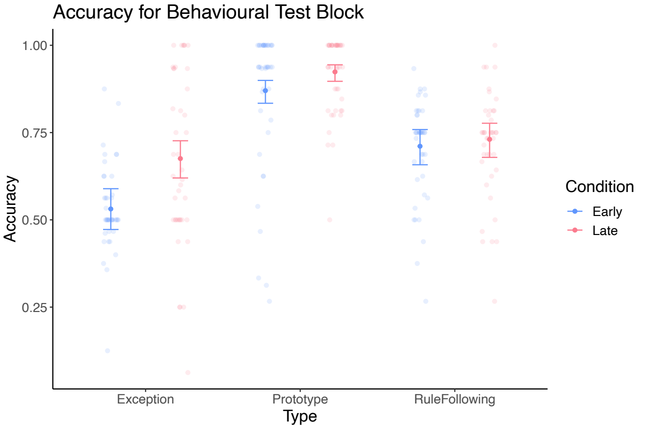

Figure 3. Categorization accuracy in the test block. Error bars indicate standard error, and translucent dots indicate participant averages.

Model Simulations and Analysis

Methods

Overview of Model Architecture To further study the impact of sequence on category learning, the rule-plus-exception task described above was simulated using a neural network model of hippocampus. This model was originally developed by Ketz, Morkonda, & O'Reilly (2013) to model how hippocampal subfields and white matter pathways support episodic memory and was later adapted by Schapiro et al. (2017) to study how hippocampus supports rapid statistical learning. In the present study, the model published by Schapiro et al. (2017) was run using Emergent7 v. 8.5.2 (Aisa, Mingus, & O'Reilly, 2008). A simplified explanation of the model's architecture is as follows: input patterns in the form of numerical arrays are presented to the model via its input layer, EC_in (which represents superficial layers of EC). During training, the model learns to replicate the pattern presented to EC_in in its output layer, EC_out (which represents deep layers of EC). The model accomplishes this goal by adjusting the weights of connections between its hidden layers, which represent DG, CA3, and CA1. Weights are updated using a combination of Hebbian and error-driven learning that mimics hippocampal theta oscillation (Ketz et al., 2013). Each layer of the model contains a grid of several units with activity levels ranging from zero to one. These units represent populations of neurons. Moreover, each layer has physiology-based properties. For example, model layers CA3 and DG have high within-layer inhibition, leading to the sparse representations characteristic of these neural subfields. Connections between layers mimic the flow of information along TSP and MSP, and the learning rate of TSP is also faster than that of MSP. The model has free parameters that allow the user to adjust the strength of connections between CA3 and CA1 and EC_in and CA1 to simulate white matter lesions. Because this study involved healthy young adults, the fully connected values from Schaprio et al. (2017) were used. The model is shown in Figure 4.

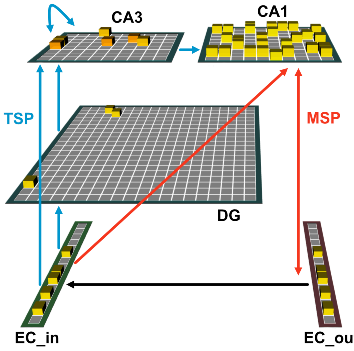

Figure 4. The neural network model used for simulations. Hidden layers represent hippocampus subfields, and inter-layer connections represent the monosynaptic and trisynaptic pathways (MSP and TSP, respectively).

Training and Testing The flower stimuli were first transformed into vectorized input patterns. Each input vector has five pairs of units, and each pair represents a feature dimension. The first four pairs (units 0 to 7) correspond to the four binary-valued dimensions, and the final pair (units 8 and 9) indicates category label. Each unit in a pair represents one of two possible values for a given dimension, so only one unit in a pair will be active (i.e., have a non-zero value) for a stimulus. For example, pointed petals may be coded as "01", round petals, by "10". In vector notation, the prototype for category A is therefore represented as "1010101010." The vector notation for each stimulus is included in Table 2.

Table 1. Vector notation for prototypes, rule-followers, and exceptions (P, R, and E, respectively). A and B indicate opposing categories. Dimensions D1–D4 match the labels in Figure 1. D5 is the category label.

Stimulus	D1	D2	D3	D4	D5
PA_1	10	10	10	10	10
PA_2	10	10	10	01	10
RA_1	10	01	10	10	10
RA_2	01	10	10	01	10
EA	10	01	01	10	10
PB_1	01	01	01	10	01
PB_2	01	01	01	01	01
RB_1	01	10	01	10	01
RB_2	01	01	10	01	01
EB	10	10	01	01	01

Two training sequences were created for the model that corresponded to the EE and LE conditions of the behavioural experiment. The number of stimuli and trial order presented to the model in each condition were identical to the sequences of the behavioural experiment, but the learning task was not separated into blocks. In a training epoch (144 trials), stimuli were presented to EC_in sequentially. With each trial, the model updated its connection weights to replicate the input pattern in its output layer. After training, the model was tested: each of the 10 stimuli were presented to the model and its settled output activity was recorded. No network weights were updated during test. The LE and EE sequences were each simulated 500 times on randomly initialized networks.

Simulation Results

The model's categorization performance was assessed by analyzing the activation of the EC_out units corresponding to category label. The cosine similarity between the activation of the two category label output units and the target category label was calculated relative to the nontarget category label using Luce's choice axiom (Luce, 1977), as defined below:

$$\text{accuracy} = \frac{\cos\,(output, target)}{\cos(output, target) + \cos\,(output, nontarget)}$$

The model accuracy closely aligned to accuracy in the behavioural test block. A general linear model with condition and type as fixed effects and batch as a random effect was fit to the data using the "stats" package in R.

As with the behavioural results, accuracy for exceptions was higher in the LE condition than in the EE condition (Figure 5; β_E = 0.047, P < .001, 95% CI [0.027, 0.067]). However, prototype accuracy was not significantly different in the LE condition (β_P = -0.004, P = .550, 95% CI [-0.018, -0.010]) and performance decreased for rule-followers in the LE condition (β_R= -0.022, P = .003, 95% CI [-0.036, -0.008]).

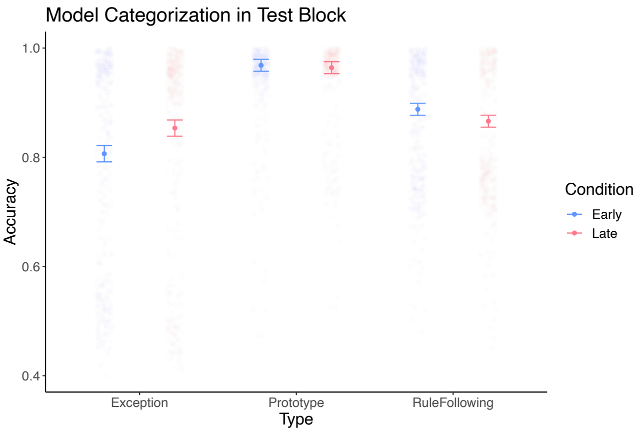

Figure 5. Model accuracy across conditions. Error bars represent standard error. Translucent dots indicate batch averages.

The model simulations capture the expected advantage for exception items in the LE condition. The model categorizes exceptions more accurately after exposure to the LE sequence compared to the EE sequence. This finding reinforces the sensitivity of the hippocampus neural network to learning sequence. However, the model also predicts significantly higher performance for rule-followers in the EE condition compared to the LE condition. A post-hoc analysis of trial-by-trial accuracy for behavioural data revealed that participant accuracy for rule-followers did drop when exceptions were introduced in the late condition. This drop maybe reflected in the model results.

Representational Similarity Analysis

We next examined how the neural network model representations varied across conditions. During the testing phase, we recorded settled activation in the hidden layers of the network corresponding to CA1, CA3, and DG. We calculated the Pearson correlations between activations for each test item (i.e., between each of the 10 stimuli) in each hidden layer. The results of this analysis are depicted in Figure 6. Note that the qualitative analysis of model results that follows is speculative in nature.

In Figure 6a, darker shades represent higher representational similarity. The overlapping representations of CA1 are reflected in the overall shade of the grids corresponding to this subregion, which are darker than those of pattern-separated CA3 and DG. The representational similarity of this region can also be visually clustered into zones: the darker lower left and upper right quadrants of the CA1 grids indicate higher intra-category similarity, whereas the lighter lower right and upper left quadrants indicate lower inter-category similarity. In other words, representations of members in the same category are more similar, whereas those of members in opposite categories are less similar. Although representations in CA3 and DG are overall more distinct (i.e., lighter) than those in CA1, higher intra- versus inter-category similarity is still evident, albeit to a lesser extent than in CA1. However, visually detecting representational differences between layers in the EE and LE conditions is quite difficult. To render these differences more apparent, Figure 6b presents the difference between representational similarity in the EE and LE conditions, as determined by subtracting the LE RSA matrix from the EE matrix for each of the three subfields.

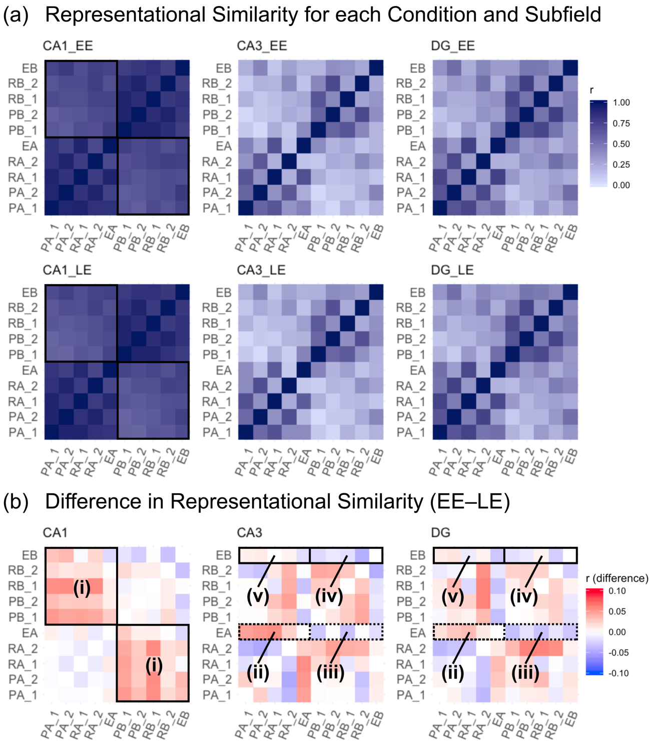

Figure 6. (a) RSA results, separated by condition and subfield. Rows and columns correspond to the 10 stimuli. Darker shades indicate higher similarity. (b) Difference in RSA results across conditions (EE minus LE). Red indicates higher similarity in the EE condition; blue, higher similarity in the LE condition.

In Figure 6b, the colour red indicates higher similarity between two stimuli in the EE condition compared to the LE condition, and blue, higher similarity between two stimuli in the LE condition compared to the EE condition. In CA1, the red colour of the upper left and lower right quadrants (zone i) denotes higher levels of inter-category similarity for stimuli in the EE condition; that is, representations of members in opposite categories are more similar in the EE condition compared to the LE condition. A higher degree of similarity between members in opposing categories indicates a blurring of category boundaries, which may reflect the model's reduced ability to distinguish exceptions, thus lowering exception categorization performance in the EE condition.

Results in CA3 and DG are slightly less pronounced but are highlighted by zones ii–v in Figure 6b. Notably, the effect on the two exceptions across conditions does not seem to be consistent. Exception A (EA; outlined in dotted black) is more similar to members of its own category (red zone ii) and less similar to members of the opposite category (blue zone

iii) in the EE condition compared to the LE condition, which should improve categorization performance in the EE condition. However, the opposite is true of Exception B (EB; outlined in solid black), which is more similar to its own category members (blue zone iv) and less similar to opposing category members (red zone v) in the LE condition than in the EE condition, which should improve performance in the LE condition. It appears that the advantage for exceptions seen in the LE condition is driven by EB alone. A post-hoc analysis of behavioural data supports this finding: when the mixed-effects logistic regression model used to analyze performance for the test blocks of the behavioural experiment is altered by replacing the fixed effect "type" by "stimulus," the effect of condition is opposite for each exception: the late condition has a positive effect on categorization accuracy for EA ($\beta = 1.573$, $P < .001$, 95% CI [1.136, 2.019]), but a negative effect for EB ($\beta = -0.455$, $P = 0.047$, 95% CI [-0.909, -0.004]). To fully understand the impacts of our manipulation, further exploration of the nature of exceptions is warranted.

General Discussion

Our aim was to explore how learning sequence impacts categorization performance. We provided behavioural evidence that delaying the introduction of exceptions significantly improves participants' ability to categorize these items. We also used a model of hippocampus to provide novel computational evidence of this area's sensitivity to trial order. Notably, this model replicated behavioural results for exceptions items, despite its lack of any direct attentional mechanism. Moreover, conducting an RSA on model representations yielded subtle but important differences in representations across conditions. In CA1, there were higher levels of inter-category representational similarity in the EE compared to the LE condition. As the degree of interleaving in each condition was equal, Carvalho and Goldstone's (2015) SAT alone cannot explain these findings.

Differences in exception representation were also seen in CA3 and DG. However, the two exception items exhibited opposite effects: one was more similar to its own category members and less similar to opposite category members in the LE condition, which one would expect to improve categorization performance in the LE condition, and the other was instead more similar to like category members and less similar to opposing category members in the EE condition, which should improve performance in the EE condition. This discrepancy was reflected in a post-hoc analysis of behavioural results: when performance for the two exception stimuli was analyzed separately, the behavioural data was consistent with RSA predictions. Asymmetries in category structure may have led to this result: when the non-diagnostic dimension is considered, EA is closer to its category rule-followers than EB. Past work has indicated that memory for exception items increases as the violated rule grows more salient (Sakamoto & Love, 2004), and this effect may have translated into categorization advantages that varied between exceptions. Because EB was more exceptional with respect to the rule-followers in its category, the impact of the manipulation may have been stronger. Future behavioural and model work should explore how adjusting the non-diagnostic dimension impacts behavioural and model results. Moreover, an attentional mechanism could be incorporated into the model to further test the impact of feature salience on exception categorization. An fMRI study of this task would allow us to test the predictions made by the neural network model and to explore how hippocampus works in concert with other brain regions during category learning problems.

Overall, this work demonstrates that performance on a rule-plus-exception category learning task can be modulated by manipulating learning sequence. We also provide novel computational evidence of hippocampus's sensitivity to this manipulation and use RSA analysis to explore the impact of trial sequence on inter- and intra-category representational similarity. The experiments presented here serve as a starting point for future studies to further explore how hippocampus and its white matter pathways are implicated in category learning tasks and in cognition more broadly.

Acknowledgements

The authors would like to thank Meg Schlichting for her suggestions and insight and members of the Mack Lab for their assistance with data collection. This research is supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2017-06753) to MLM, NSERC CGS-M Grant to EMH, Canada Foundation for Innovation and Ontario Research Fund (36601) to MLM, and the Brain Canada Future Leaders in Canadian Brain Research Grant to MLM.

References

Aisa, B., Mingus, B., & O'Reilly, R. (2008). The Emergent neural modeling system. *Neural Networks*, *21*, 1146–1152.

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, Vol. 6, p. 505. Frontiers Research Foundation.

Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*, 260–273.

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*, 821–839.

Duncan, K. D., & Schlichting, M. L. (2018). Hippocampal representations as a function of time, subregion, and brain state. *Neurobiology of Learning and Memory*, *153*, 40–56.

Goodman, G. S. (1980). Picture memory: How the action schema affects retention. *Cognitive Psychology*, *12*, 473–495.

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta Coordinated Error-Driven Learning in the Hippocampus. *PLoS Computational Biology*, *9*, e1003067.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective and Behavioral Neuroscience*, *7*, 90–108.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*, 215–233.

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 of the United States of America*, *113*, 13203–13208.

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*, 1–11.

Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, *16*, 1050–1057.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*, 311–345.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition Memory for Exceptions to the Category Rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 548–568.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*, 534–553.

Sakamoto, Y., Matsuka, T., & Love, B. C. (2004). Dimension-Wide vs. Exemplar-Specific Attention in Category Learning and Recognition. *Proceedings of the 6th International Conference of Cognitive Modeling*, 261–266. Mahwah, NJ: Lawrence Erlbaum Associates Publisher.

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, *9*, 1–11.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*. https://doi.org/10.1098/rstb.2016.0049

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*, 1–42.

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, *17*, 129–144.

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, *18*, 299–342.

Yamaguchi, S., Hale, L. A., D'Esposito, M., & Knight, R. T. (2004). Rapid prefrontal-hippocampal habituation to novel events. *Journal of Neuroscience*, *24*, 5356–5363.

Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., van Kesteren, M. T. R., & Wutz, A. (2019). Brain Mechanisms of Concept Learning. *The Journal of Neuroscience*, *39*, 8259–8266.