An *In Silico* Exploration of the Effect of Surprising Information on Hippocampal Representations

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

Category learning is our ability to generalize across experiences and apply existing knowledge to new situations. Many real-world categories adhere to a "rule-plus-exceptions" structure, wherein most items are rule-followers, but a subset of "exceptions" violate category rules. Rule-plus-exception learning seems tightly coupled with hippocampal function. Though past work has demonstrated that prediction error drives hippocampus to form distinct representations of exceptions, limited work has investigated how this process impacts existing rule-follower representations. Here we use a neural network model of hippocampus to quantify how rule-follower representations are altered by the introduction of exceptions. By recording model representations of rule-followers before and after exceptions are introduced, we computed the shift in rule-follower representation elicited by exceptions. A rule-follower's similarity to exceptions along category-relevant, but not irrelevant, dimensions predicted its degree of representational shift. This work furthers our understanding of how hippocampus supports the integration of surprising information in dynamic environments.

Keywords: hippocampus; category learning; rule-plus-exceptions; computational modelling

Introduction

Category learning is our ability to generalize across episodes, and it allows us to make inferences about the properties of novel exemplars. The utility of category learning is evident when one explores the alternative. Consider a student studying for an upcoming biology exam. Without the ability to categorize, they would have to exhaustively memorize individual properties of each cell type in the human body, like whether each cell contains a nucleus. To expedite their studying, the student can instead group living cells together and infer that cells in this category share common properties. However, many naturalistic categories contain exceptions to the rule. Mature red blood cells, for example, are living cells that do not contain a nucleus. Such a "rule-plus-exceptions" category structure requires the learner to generalize across experiences to correctly categorize rule-followers while also detecting and remembering exceptions. However, the extent to which exceptional information impacts previously learnt information is relatively unstudied. For example, when a student learns that red blood cells do not contain a nucleus, might this new information interfere with their existing knowledge of white blood cells? Here we use a neural network model of hippocampus (HC) to explore this question.

Category learning is a complex process that recruits myriad brain regions (Zeithamova et al., 2019), but an area of the brain frequently implicated in category learning is the hippocampus (e.g., Bowman & Zeithamova, 2018; Davis et al., 2012b, 2012a; Mack et al., 2016; Schapiro et al., 2018). Our long-standing understanding of HC's role in episodic memory dates to seminal studies on patient H.M. (Scoville & Milner, 1957), but more recent findings suggest that HC can also generalize across episodes (Schapiro et al., 2017; Schlichting et al., 2015). The functionally distinct white matter pathways of HC and their associated subfields may support these seemingly divergent abilities: the monosynaptic pathway traverses cornu ammonis 1 (CA1), which employs overlapping representations ideal for extracting regularities; conversely, the trisynaptic pathway includes dentate gyrus (DG) and cornu ammonis 3 (CA3) and has been associated with the encoding of exceptions (Schapiro et al., 2017; Schlichting et al., 2021). Given HC's evidenced ability to both encode distinct episodes and generalize across experiences, it seems well-suited for the demands of rule-plus-exception learning. HC also plays an established role in spatial navigation and encodes maps of the physical world that capture the spatial relationships between objects (O'Keefe & Nadel, 1978), but this function is not limited to physical space – HC can encode mappings of non-spatial structured environments such as those found in many category learning tasks. Work by Theves et al. (2019) has indicated that HC encodes distances in conceptual space; moreover, this mapping of category-relevant dimensions seems specific to category-relevant dimensions (denoted by the authors as concept space), but not irrelevant dimensions (feature space; Theves et al., 2020). HC seems especially important for encoding the conjunction of multiple category-relevant features, as is required in rule-plus-exception learning (Love & Gureckis, 2007).

The conjunctive representations attributed to HC are particularly crucial when learning to correctly categorize exceptions, and computational modelling work has elucidated the neural underpinnings of such memory structures. SUSTAIN, a prominent model of category learning, posits that category exemplars are stored in clusters that group together similar objects of the same category (Love et al., 2004). When the model encounters an item that does not fit into an existing cluster, it creates a unique cluster

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

for this item. SUSTAIN therefore tends to group rule-followers together and stores exceptions separately. Davis et al. (2012a) found that HC activation tracked SUSTAIN-derived measures of item recognition and error correction; importantly, activation was distinct for rule-followers compared to exception items. The authors concluded that specialized conjunctive representations are formed and recruited in HC to encode exceptions in rule-plus-exception learning. Such specialized representations could account for memory advantages often reported for exceptions and schema-violating information (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2006; von Restorff, 1933).

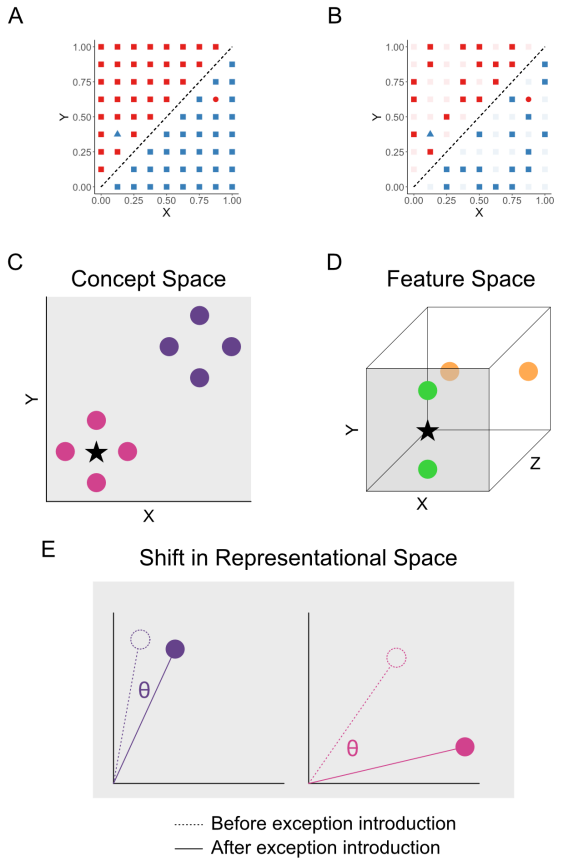

Figure 1: The category structure used for simulations. A: Category membership is defined by two dimensions, X and Y, with the category boundary along the line X = Y. Each category has an exception (EA and EB are indicated by the red circle and blue triangle, respectively). This two-dimensional space is referred to as the concept space. B: The feature space contains a third, non-diagnostic dimension, indicated by the opacity of points. C: Stimuli can be close to (pink) or far from (purple) the exception in concept space. D: Stimuli may be close to the exception in concept and feature space (green) or close in concept but far in feature space (orange). E: The proximity of an exemplar to an exception should influence the degree to which it shifts, θ; close exemplars (pink) should shift more than distal exemplars (purple).

Although exception representation in the context of category learning has been studied in relative detail, limited work has focused on how exception learning impacts rule-followers. Research on pigeon category learning suggests that the presence of cross-over exceptions, which share features of the opposing category, impedes rule-follower learning; the presence of oddball exceptions, which have an entirely unique set of features, does not (Castro et al., 2021). Moreover, Heffernan et al. (2021) found that when exceptions are introduced after extensive exposure to rule-followers (as opposed to at the outset of learning) in rule-plus-exception categorization, one's ability to detect and correctly categorize exceptions improves. However, this delayed introduction of exceptions was qualitatively associated with a transient decrease in categorization performance for rule-followers in both human participants and a computational model of HC. These findings suggest that the encoding of exceptions might come at a cost to rule-follower accuracy, especially when exceptions share properties of rule-followers. This prediction aligns with work beyond the domain of category learning: Sinclair et al., (2021) introduced prediction error while participants viewed videos and found that the ensuant surprise triggered memory updating in HC; further, videos that were more semantically similar to others in the stimulus set were more susceptible to the creation of false memories. Evidently, existing representations can be impacted by the presentation of surprising information, and the extent of this impact may be governed by an item's similarity to this surprising information.

Here we quantify the effect of surprising information on rule-follower representation using a computational modelling approach. We use a neural network model of HC and its subfields that has successfully accounted for behavioural results including episodic memory, statistical learning, and rule-plus-exception learning (Heffernan et al., 2021; Ketz et al., 2013; Schapiro et al., 2017). By recording the model's representation of rule-following stimuli in a category learning task before and after exceptions are introduced, we explore the extent to which rule-follower representations shift in model-defined hippocampal subfields and whether this shift is modulated by a rule-follower's similarity to exceptions. Moreover, by defining a category structure that contains both diagnostic and non-diagnostic dimensions, we examine how this shift is related to a rule-follower's similarity in both concept space (defined by diagnostic features) and feature space (defined by both diagnostic and non-diagnostic features; Theves et al., 2020). We predict that proximity to exceptions in concept space should be positively associated with greater shifts in rule-follower representations when exceptions are introduced (Figure 1E); however, if the model is encoding concept space rather than feature space, proximity defined along the non-diagnostic dimension should not better predict shift. We explore this effect in three hippocampal subfields, CA1, CA3, and DG.

Methods

Category Structure

A simple two-dimensional category structure was used for simulations (Figure 1A). Stimuli are defined along two continuous dimensions, X and Y, which range from 0 to 1 (inclusive) by increments of 0.125. The stimuli are divided into two categories, with the category boundary defined along the line Y = X; stimuli above this line belong to Category A, and stimuli below this line, to Category B. Stimuli that fall along the category boundary were excluded, so each category has 36 members. This category structure also contains two exceptions, EA and EB, which are respectively indicated by the red circle and blue triangle in Figure 1A.

A third feature with no bearing on category assignment was also included in the category structure (Figure 1B). This non-diagnostic feature assumes values of 0 or 1 and varies pseudo-randomly across all stimuli. Both exceptions have a non-diagnostic feature value of 1 and are surrounded by an equal number of stimuli with non-diagnostic feature values of 0 and 1 to prevent any unintentional impact of this feature on the symmetry of the category structure. Each stimulus can be expressed as an array of four values corresponding to X, Y, the non-diagnostic dimension nd, and the category label c: [x; y; nd; c]; with one-hot feature encoding, EA is therefore written as [0.875, 0.125; 0.625, 0.375; 1, 0; 0, 1], and EB, [0.125, 0.875; 0.375, 0.625; 1, 0; 1, 0].

Computational Model and Simulations

Model Architecture A computational model of HC and its subfields was used for simulations (Figure 2). This model was selected for use in the present study because, though it was originally designed to mimic episodic memory in HC (Ketz et al., 2013), it has also successfully accounted for a range of human behaviour (Heffernan et al., 2021; Schapiro et al., 2017). It is also unique compared to many prominent models of category learning in that it can accept continuously valued features as inputs, rather than discrete binary features that can only assume values of 0 or 1. The version of the model developed and made available by Schapiro et al. (2017) has been used in the present study. A thorough explanation of this model can be found in related work, but a brief overview is presented here.

This model accepts inputs at its input layer, EC_in, which represents superficial layers of entorhinal cortex. From EC_in, information flows along two pathways that mimic hippocampal white matter pathways. The monosynaptic pathway flows directly from EC_in to CA1, whereas the trisynaptic pathway traverses DG and CA3 and then CA1. CA1 is connected to the output layer, EC_out, which represents deep layers of entorhinal cortex. Note that Figure 2 depicts 15 units in EC_in and EC_out, but in the current simulations these layers were comprised of only eight units. Connections between the output and input layer simulate big loop recurrence in hippocampus. Moreover, the hidden layers corresponding to each subfield have properties that reflect their respective subfields; for example, high within-layer inhibition in DG leads to sparse activation of units within each layer, low within-layer inhibition leads to overlapping representations in CA1, and recurrent within-layer connections simulate pattern completion in CA3. The model acts as an autoencoder and tries to replicate patterns presented to EC_in in its output layer, EC_out. It does so by adjusting weights between its hidden layers through Hebbian learning that mimics hippocampal theta oscillations. The learning rate along TSP is also faster than MSP. Although connections between layers can be reduced to simulate lesions, a fully connected version of this model are used for the present study.

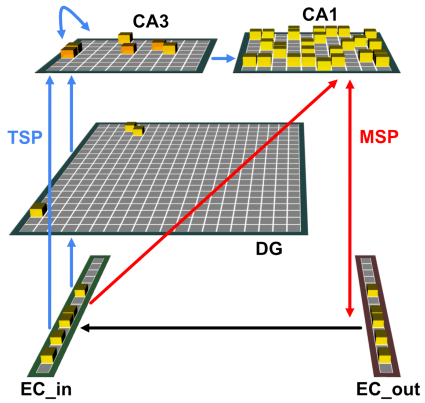

Figure 2: The neural network model of hippocampus (HC) used in the present study. The model attempts to reproduce inputs presented to its input layer (EC_in) on its output layer (EC_out). It does so by adjusting weights between its hidden layers (DG, CA1, and CA3). Information flow mimcs the hippocampal trisynaptic (TSP) and monosynaptic (MSP) pathways.

Simulations The aim of these simulations was to explore how exception introduction shifts rule-follower representations. Separate simulations were run for EA and EB to isolate the effects of each exception. These simulations can be separated into two distinct training and testing phases: pre- and post-exception. In the pre-exceptions phase, all rule-following stimuli were first presented once to the model in a training epoch. Training epochs were identical for both EA and EB simulations. Following the training epoch, the model's response to each rule-follower was tested and settled activation in each of the model's hidden layers was recorded. In the post-exceptions phase, this trained model was presented with either EA or EB in a single-trial epoch. Activation was then immediately recorded in a transient, post-exceptions testing phase that again tested settled activation for each rule-follower. The simulations for the two exceptions were both run 500 times (batches) with random initializations of model weights.

Quantifying Shift Representational similarity analysis (RSA) was conducted by computing the Pearson correlation between the settled activation of each stimulus in each of the

model's hidden layers. The pre-exception RSA matrices were compared to post-exception matrices computed using the settled activation immediately after each exception was introduced. Each row of the RSA matrices can be considered a vector that reflects a stimulus's location in a 70-dimensional similarity space. Shift θ_R for rule-follower R was formalized as the angle between the vectors reflecting a rule-follower's pre- and post-exception representation and was computed as follows:

$$\theta_R = \cos^{-1}\frac{a \cdot b}{|a||b|}, \quad (3)$$

where a and b are the vectors corresponding to the pre- and post-exception representation of a rule-follower. Shift was calculated for each rule-follower in each of the model's hidden layers (corresponding to subfields CA1, CA3, and DG). Formalizing shift in this manner allowed us to get a measure of an exception's influence on each individual rule-follower and to consequently explore whether this influence varied as a function of distance. Distance was quantified as either a discrete or continuous value. In the discrete analysis, rule-followers adjacent to exceptions were labelled "close" and rule-followers opposite the category space were labelled "far" from exceptions. Paired t-tests were used to explore differences between these groups in each subfield for both EA and EB. The continuous distance between a stimulus and an exception was quantified as Euclidean distance in both concept space, which included only category-relevant dimensions, and feature space, which also included the non-diagnostic feature. A general linear mixed-effects model with distance as a fixed effect and simulation number (batch) as a random effect was used to explore the relationship between shift and distance:

$$\theta_R \sim D_R + (1|Batch), \quad (4)$$

where D_R is the distance in either concept or feature space.

Results

Our aim was to explore how the introduction of exceptions differentially affects representations of rule-followers that are close to and far from exceptions. We also explored whether this relationship varied between concept and feature space. To that end, we first exposed the model to the 70 rule-following stimuli and recorded settled activation at test. For each batch and each hidden layer, we then used Pearson correlation coefficients to derive RSA matrices.

After the model was exposed to EA or EB, we compared the pre-exceptions RSA matrix to the post-exceptions RSA matrix. Each 70-unit row in these matrices reflects a rule-follower stimulus's similarity to all other rule-following stimuli. Equation 3 was used to compute the angle between pre- and post-exception representation for each rule-follower in each of the 500 batches and in each subfield (CA1, CA3, and DG).

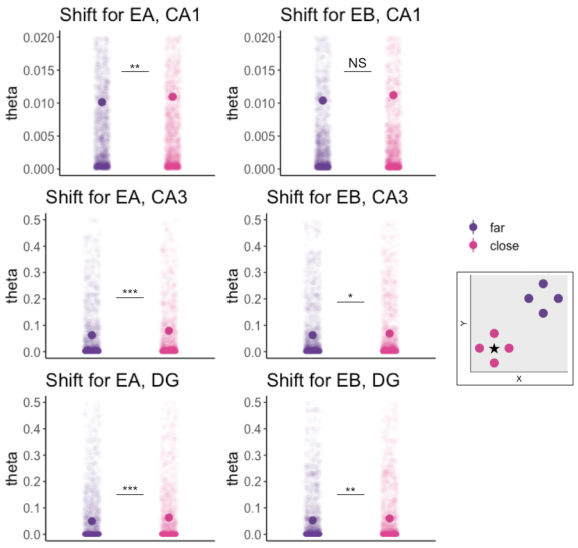

Figure 3: Shift experienced by items far from exceptions (pink) compared to shift experienced by items close to exceptions (purple). Lighter points indicate shift during individual batches; opaque points indicate batch averages. Error bars reflect standard error. Items closer to exceptions shifted more (with the exception of items in CA1 following the introduction of EB).

We compared shift in rule-followers adjacent to exceptions to shift in stimuli located further from exceptions (pink and purple groups in the inset of Figure 3) using paired t-tests. The effect of EA and EB was explored separately in each subfield of the hippocampal model. Following the introduction of EA, average shift for close rule-followers (M_{close} = 0.0110, SD_{close} = 0.0199) in CA1 was significantly higher than average shift for rule-followers far from exceptions (M_{far} = 0.0101, SD_{far} = 0.0168; $t(499) = 2.6362$, P = .009). The same pattern held in layers CA3 (M_{close} = 0.0792, SD_{close} = 0.1939; M_{far} = 0.0624, $SD_{far.}$= 0.1317; $t(499) = 4.406$, P < .001) and DG (M_{close} = 0.0632, SD_{close} = 0.1485; M_{far} = 0.0498, $SD_{far.}$= 0.1072; $t(499) = 4.6334$, P < .001). Following the introduction of EB, average shift in CA3 was higher for rule-followers close to versus far from exceptions (M_{close} = 0.0690, SD_{close} = 0.1588; M_{far} = 0.0624, $SD_{far.}$= 0.1308; $t(499) = 2.017$, P = 0.044); the same was true in DG (M_{close} = 0.0599, SD_{close} = 0.1312; M_{far} = 0.0529, $SD_{far.}$= 0.1107; $t(499) = 2.7992$, P = .005). No significant difference was observed between close and far rule-followers following the introduction of EB in CA1 (M_{close} = 0.0112, SD_{close} = 0.0233; M_{far} = 0.0104, $SD_{far.}$= 0.0193; $t(499) = 1.5291$, P = .1269).

We next explored whether our simulations would support existing findings that HC maps concept space. If distance in feature space did not influence shift, we predicted that there would be no difference in shift for rule-followers close to an exception in both concept and feature space (green items in the inset of Figure 4) and items close to an exception in concept but not feature space (orange items in the inset of

Figure 4). This prediction held true; paired t-tests indicated that there were no differences in shift between stimuli in these groups following the introduction of EA and EB (all P values > 0.05; Figure 4).

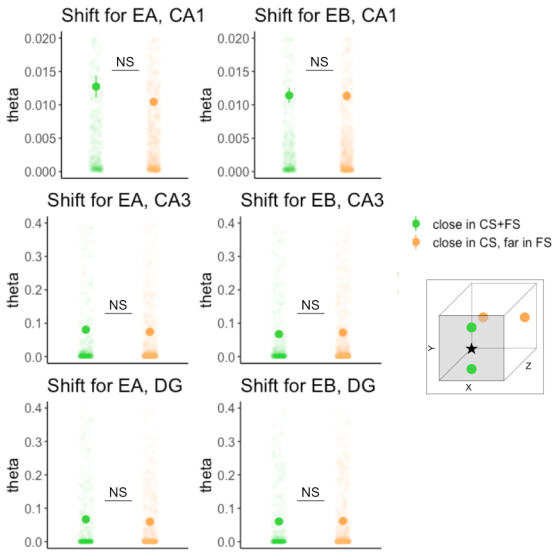

Figure 4: Shift experienced by items close in both concept space (CS) and feature space (FS; green) and items close in CS but not FS. Error bars indicate standard error. No significant difference was observed across these groups in any layer, suggesting that the HC model was encoding concept space.

Having confirmed that proximity to exceptions in concept space did predict higher shift, we next explored to what extent a rule-follower's distance from an exception in concept and feature space predicted shift θ_R when distance was treated as a continuous value; this was again examined separately following the introduction of EA and EB using the general linear mixed-effects model defined in Equation 4. We used the Akaike information criterion (AIC) to compare these models to see if distance in concept space better predicted shift than distance in feature space. The results from these model fits are depicted in Figure 5 and are described quantitatively below.

In layer CA1, after the introduction of EA, rule-followers further from EA shifted less as distance increased when distance was measured in both concept and feature space (β = -0.0008, P < .001, 95% CI [-0.0012, -0.0003] and β = -0.0005, P = .001, 95% CI [-0.0008, -0.0002], respectively). The concept space model (AIC = -109,312.1) provided a better fit than the feature space model (AIC = -109,309.8). The results in CA1 following the introduction of EB were similar. However, following the introduction of EB, distance in concept but not feature space significantly predicted shift (β = -0.0012, P < .001, 95% CI [-0.0018, -0.0007] and β = -0.0001, P = .596, 95% CI [-0.0005, 0.0003], respectively). The concept space model (AIC = -99,762.5) provided a better fit than the feature space model (AIC = -99,745.0).

In layer CA3, after the introduction of EA, lower distance from EA in both feature space and concept space significantly predicted higher shift (β = -0.0145, P < .001, 95% CI [-0.0188, -0.0102] and β = -0.0060, P < .001, 95% CI [-0.0088, -0.0033], respectively). The AIC comparison indicated that the feature space model (AIC = -30,296.9) provided a better fit than the concept space model (AIC = -30,270.7). However, following the introduction of EB, lower distance in concept space significantly predicted greater shift (β = -0.0068, P = .011, 95% CI [-0.0114, -0.0023]), but the opposite was true in feature space, where greater distance predicted greater shift (β = 0.0031, P = .047, 95% CI [0.0000, 0.0061]). Still, the concept space model (AIC = -28,614.2) provided a better fit than the feature space mode (AIC = -28,608.5).

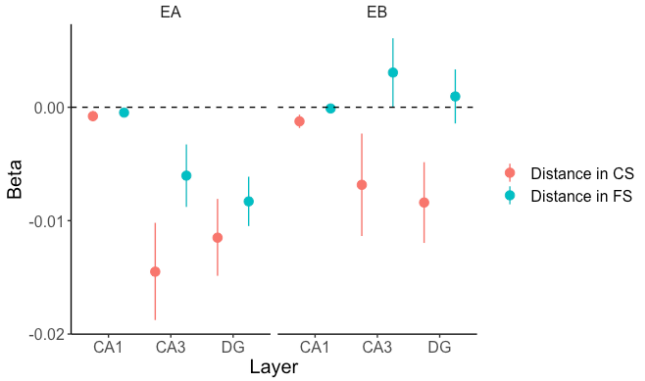

Figure 5: The relationship, beta, between distance and shift following the introduction of each exception in each layer. Red points reflect the relationship between shift and distance in concept space; blue, the relationship between shift and distance in feature space. A negative beta value indicates that lower distance was associated with greater shift. Error bars reflect 95% confidence intervals.

In layer DG, distance in both concept and feature space was predictive of shift following the introduction of EA (β = -0.0115, P < .001, 95% CI [-0.0149, -0.0081] and β = -0.0083, P < .001, 95% CI [-0.0105, -0.0061], respectively), but the feature space model (AIC = -38,377.0) had a better fit than the concept space model (AIC = -38,366.1). Shift following the introduction of EB was predicted by distance in concept space (β = -0.0084, P < .001, 95% CI [-0.0120, -0.0048]) but not feature space (β = 0.0010, P < .432, 95% CI [-0.0014, 0.0033]), with the former (AIC = -36,779.0) providing better fit than the latter (AIC = -36,757.4).

Discussion

We sought to explore how the introduction of exceptions impacts existing hippocampal representations of category rule-following stimuli. If similarity to surprising information affected hippocampal representation (as formalized by the neural network model), we expected rule-followers closest to exceptions to experience greater shift than those far from exceptions. To explore whether a computational model of HC maps concept or feature space, we also compared how a rule-

follower's shift was predicted by its distance from exceptions as calculated using either all features (feature space) or only conceptually relevant features (concept space). If the hippocampal model does map concept space, we expected distance in feature space should offer no improvement in predicting shift. These analyses were conducted with distance as both a discrete and continuous value.

Our prediction that a rule-follower's distance from exceptions would be inversely proportional to its shift was supported in both discrete space (items close to exceptions shifted more than items far from exceptions) and in continuous space (distance from an exception was a significant predictor of the degree of shift an item experienced). These results provide compelling evidence that the introduction of exceptions may have targeted impacts on one's existing knowledge. However, these findings are specific to the effect of exceptions on novel information. Consolidated information stored in cortex is likely to be impervious to such shifting effects; indeed, recent modelling work has indicated that hippocampus and neocortex collaborate during sleep to protect existing memories while integrating novel information (Singh et al., 2022).

Both our discrete and continuous analyses further indicated that the distance between a previously encountered rule-follower and a novel exception significantly predicts representational shift in the three studied hippocampal subfields. These results support existing evidence that HC maps objects according to category-relevant, but not non-diagnostic, dimensions (Mack et al., 2016; Theves et al., 2020). Though the selected hippocampal model has successfully replicated several behavioural findings, it simply instantiates known properties of HC and its computations and acts as a simple autoencoder. As no brain region acts in isolation, the present findings offer only an abstract prediction of how hippocampal representations might be impacted by surprising information. Behavioural and fMRI studies supporting these results are necessary.

Though our findings generally supported our hypotheses, there were some unexpected results. In our discrete analysis, though the trend of the data matched our predictions, there was no significant difference between shift of close and far items in CA1 following the introduction of EB. CA1 falls along MSP, which has a slower learning rate than TSP. In this work we only measured the transient response of the network to exception introduction. As MSP more slowly incorporates overlapping information in CA1, akin to statistical learning (Schapiro et al., 2017), is possible that stronger effects would be seen in CA1 after multiple presentations of the exceptions. A second unexpected finding occurred in DG: here, shift was better predicted by distance in feature space than concept space following the introduction of EA. DG has sparse, non-overlapping representations ideal for encoding unique features of exemplars (Sučević & Schapiro, 2022), and it is possible that this region may encode more detail to minimize overlap, regardless of category relevance. Finally, minor differences in the results across the two categories must be acknowledged. Though a tightly controlled, rather than randomized, trial sequence may lessen any differences between EA and EB, symmetry across the categories was not the intention of this experiment. Instead, the results provide further evidence of how highly sensitive hippocampal encoding functions are to the order of information.

The methods used in this work and the results obtained from the present analyses may provide a foundation for future studies on rule-plus-exception category learning. The model can be adjusted to introduce virtual lesions to the mono- and trisynaptic pathways, and an evident extension of the current work is a model-based lesioning study. Such work could expand upon evidence of the unique contributions of hippocampal white matter pathways to rule-plus-exception learning. The subfield-level predictions provided by the HC model could also be used to design neuroimaging studies that employ similar methods to those described in this work to explore the impact of surprising information on hippocampal representations *in vivo*. Recent diffusion-weighted imaging work has indicated that the integrity of the trisynaptic pathway is related to individual differences in exception learning (Schlichting et al., 2021). The model simulations described in the present work could be used in conjunction with neuroimaging and behavioural studies to explore how the unique configuration of both the mono- and trisynaptic pathway support individual differences in the integration of surprising information. Experimental data will further serve to validate predictions made by the present model.

A final contribution of this work is its potential to inspire updates to well-established models of category learning. Though cluster-based models of category learning like SUSTAIN (Love et al., 2004) have successfully accounted for behaviour across a wide range of category learning tasks, these models do not adjust clusters corresponding to existing information when surprising new information (i.e., an exception) is encountered. The findings in the present study indicate that surprising information elicits updates to existing representations in addition to the formation of new representations. Adjustments to existing models of category learning may be warranted to more robustly capture category learning behaviour. Moreover, though human learners can consolidate new with old information via hippocampal-neocortical interaction (McClelland et al., 2020), artificial learners suffer from "catastrophic forgetting" – that is, they tend to forget what they have learned from one task when they switch to a second task, especially when tasks are of intermediate similarity (Lee et al., 2021). A better understanding of encoding processes may help researchers design networks more robust to such issues.

Overall, this study offers a novel computational approach to quantify how similarity to surprising information impacts shifts in existing representations. Distance from previously encountered rule-following stimuli in concept and feature space can predict shift in various hippocampal subfields. We live in a dynamic, ever-changing world, and these findings shed light on how we may accommodate new information at the expense of what we have already learned.

Acknowledgements

The authors would like to thank Morgan Barense and Meg Schlichting for their suggestions and insight. Thank you also to members of Mack Lab and Budding Minds Lab for their feedback and support. This research is supported by the Canadian Institutes of Health Research (PJT-178337) Grant to MLM, the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2017-06753) to MLM, the NSERC CGS-D Grant to EMH, the 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

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10). https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Castro, L., Yang, S., Savic, O., Sloutsky, V., & Wasserman, E. (2021). Not all exceptions are created equal: Learning of exceptions in pigeons' categorization. *Psychonomic Bulletin & Review*, *28*(4), 1344–1353. https://doi.org/10.3758/s13423-021-01912-1

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*(2), 260–273. https://doi.org/10.1093/cercor/bhr036

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*(4), 821–839. https://doi.org/10.1037/a0027865

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. https://doi.org/10.1038/s41598-021-00864-9

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta Coordinated Error-Driven Learning in the Hippocampus. *PLoS Computational Biology*, *9*(6), e1003067. https://doi.org/10.1371/journal.pcbi.1003067

Lee, S., Goldt, S., & Saxe, A. (2021). Continual Learning in the Teacher-Student Setup: Impact of Task Similarity. *ArXiv*. http://arxiv.org/abs/2107.04384

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 90–108. https://doi.org/10.3758/CABN.7.2.90

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/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 of the United States of America*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

McClelland, J. L., McNaughton, B. L., & Lampinen, A. K. (2020). Integration of new information in memory: New insights from a complementary learning systems perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1799), 20190637. https://doi.org/10.1098/rstb.2019.0637

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press ; Oxford University Press.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition Memory for Exceptions to the Category Rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 548–568. https://doi.org/10.1037/0278-7393.21.3.548

Sakamoto, Y., & Love, B. C. (2006). Vancouver, Toronto, Montreal, Austin: Enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin and Review*, *13*(3), 474–479. https://doi.org/10.3758/BF03193872

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), 1–11. https://doi.org/10.1038/s41467-018-06213-1

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*(1711). https://doi.org/10.1098/rstb.2016.0049

Schlichting, M. L., Gumus, M., Zhu, T., & Mack, M. L. (2021). The structure of hippocampal circuitry relates to rapid category learning in humans. *Hippocampus*, *31*(11), 1179–1190. https://doi.org/10.1002/hipo.23382

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9151

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21. https://doi.org/10.1136/jnnp.20.1.11

Sinclair, A. H., Manalili, G. M., Brunec, I. K., Adcock, R. A., & Barense, M. D. (2021). Prediction errors disrupt hippocampal representations and update episodic

memories. *Proceedings of the National Academy of Sciences*, *118*(51), e2117625118. https://doi.org/10.1073/pnas.2117625118

Singh, D., Norman, K. A., & Schapiro, A. C. (2022). *A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation bioRxiv*. https://doi.org/10.1101/2022.01.31.478475

Sučević, J., & Schapiro, A. C. (2022). *A neural network model of hippocampal contributions to category learning* (p. 2022.01.12.476051). bioRxiv. https://doi.org/10.1101/2022.01.12.476051

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. *Current Biology*, *29*(7), 1226-1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Theves, S., Fernández, G., & Doeller, C. F. (2020). The Hippocampus Maps Concept Space, Not Feature Space. *Journal of Neuroscience*, *40*(38), 7318–7325. https://doi.org/10.1523/JNEUROSCI.0494-20.2020

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, *18*(1), 299–342. https://doi.org/10.1007/BF02409636

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*(42), 8259–8266. https://doi.org/10.1523/JNEUROSCI.1166-19.2019