Compositional Neural Representations in the Hippocampal Formation and Prefrontal Cortex Underlie Visual Construction and Planning

Philipp Schwartenbeck (pschwartenbeck@gmail.com)
Wellcome Trust Centre for Neuroimaging, Queen Square 12
London, WC1N 3BG United Kingdom

Alon Baram (alonbaram2@gmail.com)
Wellcome Centre for Integrative Neuroimaging, University of Oxford,
John Radcliffe Hospital, Oxford OX3 9DU, UK

Shirley Mark (markshir@gmail.com)
Wellcome Trust Centre for Neuroimaging, Queen Square 12
London, WC1N 3BG United Kingdom

Zeb Kurth-Nelson (zebkurthnelson@gmail.com)
DeepMind, London, UK

Raymond Dolan (r.dolan@ucl.ac.uk)
Max Planck University College London Centre for Computational Psychiatry and Ageing Research, 10-12 Russell Square
London, WC1B 5EH United Kingdom

Tim Behrens (behrens@fmrib.ox.ac.uk)
Wellcome Centre for Integrative Neuroimaging, University of Oxford,
John Radcliffe Hospital, Oxford OX3 9DU, UK

Abstract:

The hippocampal formation is critical for spatial and relational inference in navigation problems. The neural code underlying such inference is *factorized* in the entorhinal cortex (EC) and *conjunctive* in the hippocampus (HC). A *factorized* code implies a separate encoding of sensory and relational knowledge, which can be flexibly *conjoined* to an object representation that reflects both sensory and relational properties. We hypothesize that the same neural mechanisms are employed in complex decision-making and compositional planning, which requires the flexible generalization of knowledge to novel instances. We tested this hypothesis in a task where subjects had to construct novel visual objects based on a set of basic visual building blocks and relations. We found behavioral evidence that subjects form a hierarchical representation of this task that allows them to flexibly apply compositional knowledge to novel stimuli. Using fMRI adaption, we found evidence that the construction of novel objects depends on compositional neural representations in HC-EC and medial prefrontal cortex (mPFC). Further, we found that these structures also encoded purely relational information, indicative of a factorized representation. These results suggest that compositional neural representations in the hippocampal formation and prefrontal cortex enable the generalization of abstract knowledge to novel stimuli during visual construction.

Keywords: hippocampal formation; cognitive map; compositional planning; combinatorial generalization

Introduction

The hippocampal formation encodes a 'cognitive map' that allows animals to navigate successfully (Behrens et al., 2018; Tolman, 1948). A cognitive map provides an efficient neural representation of knowledge about the structure of the world that enables flexible and generalizable behavior. In the context of spatial navigation, the instantiation of a cognitive map has been associated with place cells in the HC (O'keefe & Nadel, 1978) and grid cells in the EC (Hafting, Fyhn, Molden, Moser, & Moser, 2005). It has been suggested that place cells encode individual states within a task, such as a particular location in a maze, whereas grid cells encode relational information about those states, such as likely transitions between locations (Momennejad et al., 2017; Stachenfeld, Botvinick, & Gershman, 2016). Recently, the same neural architecture has been implied in non-spatial navigation based on a cognitive map of task structure (Aronov, Nevers, & Tank, 2017; Constantinescu, O'Reilly, & Behrens, 2016; Garvert, Dolan, & Behrens, 2017).

A key principle of functional organization within the hippocampal-entorhinal system is a factorized and conjunctive neural code (Behrens et al., 2018; Manns & Eichenbaum, 2006), which can be organized hierarchically (Stachenfeld et al., 2016). This implies that

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0

EC encodes a separate or *factorized* representation of sensory and relational components of a stimulus, whereas HC encodes its *conjunction* and thus forms a high-dimensional representation of a stimulus within a specific structure.

We hypothesized that the same neural mechanisms are at play during complex decision-making and flexible planning behavior. A central aspect of flexible behavior is the generalization of abstract knowledge to novel instances. For example, if we would offer you some tea jelly for dessert, there is a fair chance that you would be able to express your preference for this food item even though you have probably never tasted it before. Importantly, previous work has shown that the construction of such novel goods relies on neural representations in the HC and mPFC (Barron, Dolan, & Behrens, 2013).

In the present study, we investigated the neural representations that underlie such compositional reasoning and the generalization of abstract knowledge to novel instances, called *combinatorial generalization* (Battaglia et al., 2018). Our key hypothesis was that the generalization of knowledge critically depends on *compositional neural representations*, where basic building blocks and relational knowledge can be flexibly combined to form novel conjunctive representations (Behrens et al., 2018; Battaglia et al., 2018).

Results

To investigate the neural representations underlying compositional planning, we developed a task in which subjects learned to construct visual objects using a toolkit of building blocks and relations. As illustrated in Figure 1A, subjects were trained to combine different building blocks by putting them on top or beside each other, without worrying about the physical stability of the resulting object. This task allowed us to probe whether a compositional neural representation in terms of a cognitive map of task structure would emerge after training, based on which a given visual object can be decomposed into its constituent building blocks and relations.

To test whether compositional neural representations can be organized hierarchically, we added an additional layer to the task. During early training, subjects were repeatedly tasked to build specific visual objects that were combinations of two basic building blocks. Later on, they were tasked to build larger visual objects, which often could be decomposed into two of the smaller visual objects from early training. Thus, although never instructed explicitly, subjects were exposed to a set of 'compositional building blocks', which allowed an efficient decomposition of larger visual objects (see bottom panel of Figure 1A). Analysis of participants' behavior revealed that when they had to construct large visual objects, they relied on 'compositional building blocks' more often than predicted by chance (Figure 1B). These findings suggest that the participants indeed formed a hierarchical representation of this task.

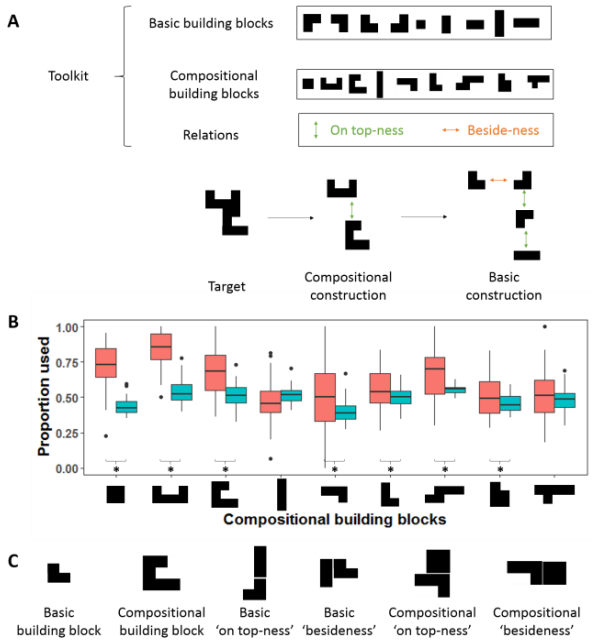

Figure 1: A) Subjects were trained to use basic and compositional building blocks and combine them via 'ontopness' and 'besideness' to construct visual objects. B) Subjects used (red) 'compositional building blocks' more often than predicted by chance (blue). C) Stimulus categories during scanning.

To probe the neural representation underlying this task, we employed an fMRI adaption paradigm after two days of behavioral training. fMRI adaption is a measure for the similarity of neural representations, based on the notion that a succession of two similar stimuli elicits higher adaption (i.e. a reduced BOLD response for the second stimulus) compared to a succession of dissimilar stimuli (Barron, Garvert, & Behrens, 2016). In consequence, one can use such 'cross-stimulus adaption' measures to probe the representational similarity of different stimuli.

In the scanner, subjects passively viewed *novel* visual objects. These objects were either a combination of two basic or two compositional building blocks connected via 'on top-ness' or 'besideness', or one basic or compositional building block alone (see Figure 1C). Participants were tasked to think about the construction of these objects. To ensure that subjects engaged in this task, ten percent of stimuli were followed by a 'catch trial', in which subjects were asked about the construction of the previous object.

First, we asked whether this visual construction task depended on neural representations in the hippocampal formation and prefrontal cortex. To test this, we

analyzed fMRI adaption for individual building blocks followed by compounds that contained these particular building block (or vice versa). The key comparison was between conditions where compound AB was preceded by building block A (or vice versa) compared to building block C. Higher adaption for AB when preceded by A compared to when preceded by C would reflect a neural representation for building blocks within a compound, and thus a compositional representation necessary for constructing a novel visual object.

This analysis revealed effects in the HC underlying this visual construction task. Specifically, we found strong *basic building block-compound* adaption in the medial temporal lobe (Figure 2A). In a region of interest analysis[1], we found a bilateral effect for basic building block-compound adaption in HC extending into EC; right: p=0.001, t_{peak}=5.95 [16 -38 -6]; p=0.003, t_{peak}=5.44 [20 -8 -12] and left: p=0.004, t_{peak}=5.26 [-26 -30 -16]; p=0.006, t_{peak}=5.13 [-24 2 -16]; p=0.031, t_{peak}=4.39 [-12 -14 -24]; p=0.044, t_{peak}=4.23 [-8 -44 8]. We also found a whole brain effect for this analysis in a large cluster containing both left and right HC, p<0.001, t_{peak}=5.95 [16 -38-6], cluster size = 8797[2].

We also found a small-volume corrected effect for *compositional building block-compound* adaption in the bilateral hippocampus; right: p=0.045, t_{peak}=4.27 [20 -30 -6] and left: p=0.005, t_{peak}=5.25 [-36 -24 -20]; p=0.036, t_{peak}=4.34 [-12 -46 2]. More exploratory whole brain analysis revealed an effect for compositional building block-compound adaption in mPFC, p<0.001, t_{peak}=6.10 [-2 24 22], cluster size = 1822 (Figure 2B). To control for adaption of basic building blocks within a compositional building block in this contrast (see bottom panel of Figure 1A), we used *basic compounds* (i.e. a combination of two basic building blocks) that contained two building blocks that were also part of the *compositional compound* as a control condition (see bottom panel of Figure 2B). However, we found the same effects when using compositional building blocks that were not part of the compound (i.e. C preceding AB) as a control condition.

We also found basic (not shown) and compositional (Figure 2B) adaptation effects in higher visual and parietal cortex.

Taken together, these results suggest that this visual construction paradigm relied on compositional neural representations in the hippocampal formation and mPFC.

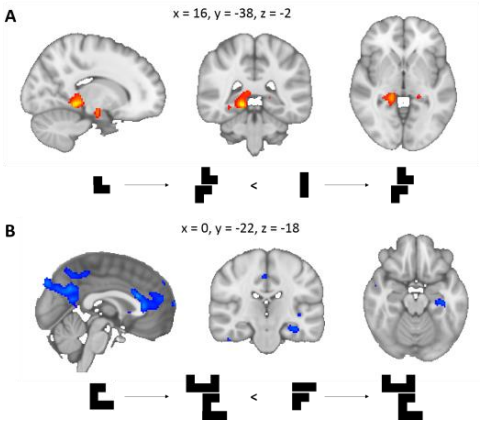

Figure 2: A) Effects for *basic building block-compound* adaption. Masked for HC and EC using a histological mask (Eickhoff et al., 2005) B) Whole-brain effects for *compositional building block-compound* adaption. Effects thresholded at T=3.5.

If the neural representation underlying flexible generalization in this task is truly compositional, this predicts a neural space that encodes objects purely in terms of their relational properties. In the simplest case, this implies representing objects in terms of their graph structure, i.e. whether two (basic or compositional) building blocks were combined with 'on-topness' or 'besideness'. Based on previous work suggesting that neurons in the hippocampal formation, particularly EC, encode relations or transitions between states (Constantinescu et al., 2016; Garvert et al., 2017), we hypothesized to find such a purely relational representation in the HC-EC system.

To test this, we probed for fMRI adaption that reflects a purely relational representation, namely contrasting same relation transitions (i.e. an object constructed with ontopness/besideness preceded by an object with the same relation) with different relation transitions. In this analysis, we found a small-volume corrected effect in left anterior EC for basic compounds, p=0.052, t_{peak}=4.17 [-16 4 -30] (Figure 3A). We did not detect an effect in right HC-EC or any other whole brain effects for *basic compound relation* adaption, or any effects for *compositional compound relation* adaptation.

Further, we tested the same question in a second analysis probing for adaption 'within' a compositional compound. As displayed in Figure 1A, *compositional compounds* are constructed with two *compositional building blocks*, which themselves consist of two *basic building blocks*. Importantly, these two compositional building blocks can be built with the same or different relations. Consequently, we expect stronger

[1] All small volume corrected analyses are based on histological masks combining the HC subiculum and EC for the left or right hemisphere, and are reported on a cluster-forming threshold of p<0.001, family-wise error corrected at the peak level.

[2] All whole brain level analyses are reported at p<0.001, family-wise error corrected on the cluster level.

(parametric) adaption for a higher proportion of 'same' relation solutions within a compositional compound. We tested this hypothesis and found a small-volume corrected effect in the HC that reflected 'within object' relational adaption; right p=0.028, t_{peak}=4.56 [24 -36 -4] and left: p=0.039, t_{peak}=4.37 [-20 -36 -4]. On a whole-brain level, we also detected parametric effects in parietal cortex, p<0.001, t_{peak}=5.90 [-12 -36 44], cluster size = 4247; callosal body, p-0.007, t_{peak}=5.65 [4 -14 24], cluster size = 348; inferior frontal gyrus, p=0.008, t_{peak}=5.07 [-36 34 6], cluster size = 333; and mPFC p=0.012, t_{peak}=4.50 [0 46 2], cluster size = 308.

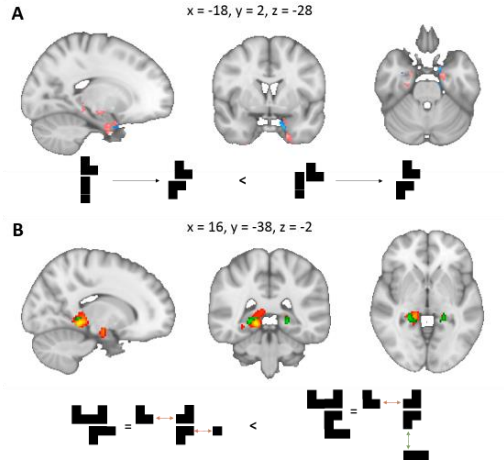

Figure 3: A) Adaption effect in anterior EC for basic 'same relation' compound transitions (light blue). Overlaid is a previously reported effect in EC for relational navigation in conceptual space (Constantinescu et al., 2016, pink). Effects thresholded at T=2.3. B) Parametric adaption effect in HC reflecting the proportion of 'same relation' solutions in compositional compounds (green), overlaid on Figure 2A. Effects thresholded at T=3.5. All effects masked for HC and EC.

Conclusion

We developed a paradigm in which subjects had to use abstract compositional knowledge to construct novel visual objects. We found evidence for such compositional neural representations in the hippocampal formation and mPFC. The latter effect was specifically pronounced for compositional building blocks within larger compounds, suggestive of a hierarchical organization in line with behavioral measures. Further, we found evidence suggesting that objects were encoded purely in terms of their relational properties in the hippocampal formation and mPFC. Taken together, our results suggest that compositional representations in the HC-EC system and mPFC underlie the flexible construction of novel stimuli, which is a central aspect of flexible decision-making and compositional planning.

Acknowledgments

We thank Avital Hahamy for very helpful comments on this manuscript. TB was supported by a Wellcome Trust grant (HMR00560.001).

References

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647), 719–722.

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, 16(10), 1492–1498.

Barron, H. C., Garvert, M. M., & Behrens, T. E. J. (2016). Repetition suppression: a means to index neural representations using BOLD? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1705), 20150355.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv*.

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-nelson, Z. (2018). What Is a Cognitive Map ? Organizing Knowledge for Flexible Behavior. *Neuron*, 100(2), 490–509.

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science* (New York, N.Y.), 352(6292), 1464–8.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, 25(4), 1325–1335.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. *eLife*, 6, 270–273.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*.

Manns, J. R., & Eichenbaum, H. (2006). Evolution of declarative memory. *Hippocampus*, 16(9), 795–808.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9), 680–692.

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

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, 20, 1643–1653.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208.