Inference in replay through factorized representations

Yunzhe Liu; Ray Dolan; Zeb Kurth-Nelson; Tim Behrens

Summary: Humans can make rich inferences from little data by generalising structural knowledge from past experience. It has been theorized such inferences rely on internal models of the world, and such world-models may be supported by the same neural mechanisms underpinning relational reasoning in space, such as hippocampal replay in rodents. In replay, patterns of cellular firing during rest spontaneously play out past, and potential future spatial trajectories. There is evidence for replay of never-before-experienced sequences that are consistent with the geometry of a spatial map. Here we ask whether, like spatial geometry, learnt arbitrary structures can impose constraints on replay. We conducted two studies in human subjects using magnetoencephalography (MEG) to measure fast spontaneous sequences of representations. In each study, we first trained participants on a rule that defined a permutation over a sequence of objects. We then presented them with a novel set of objects in a presentation order differed from the order implied by the structure. During subsequent rest, we found rapid replay of trajectories following the rule-defined as opposed to the experienced order in both studies. As in rodents, the preponderance of reverse trajectories along a sequence greatly increased after that sequence was rewarded. Within a replay event, representations of objects were preceded 50 ms by abstract factorized codes reflecting the sequence identity and the object position within a sequence. Finally, analogous to preplay, spontaneous sequences of the abstract structure representation played out even before actual experience with objects. We argue this factorized representation facilitates generalization of previously-learned structure to new objects.

Additional details: In our previous work, we have managed to detect fast sequences of decoded visual objects during either online planning (Kurth-Nelson et al. Neuron 2016), or offline rest (Liu et al. Cosyne, 2018). This method is inspired by measures in rodent spike data analysis. It used a time-lagged correlation to look for temporal relationship between decoded state reactivations and operationalizes the degree to which decoded states follow the specified transition matrix in either a forward or reverse direction. Using this method in Study 1, we have established the relationship between human non-spatial state sequences and rodent hippocampal replay, we found human offline replay occurred in sequences accelerated compared to actual experience and reversed direction after reward learning, similar as rodents. Notably, replay did not simply recapitulate visual experience, but instead followed the sequence implied by the learnt abstract knowledge, suggesting replay can generalize structure knowledge to new experience (Liu et al. Cosyne, 2018).

But how might replay facilitate such structure generalisation? During hippocampal spatial replay events, coherent replays of the same trajectories can be recorded in both medial entorhinal (mEC) and visual cortices. Whilst visual representations encode the sensory properties of a particular event, mEC representations encode structural information (such as spatial relationships), divorced from their sensory properties. One intriguing possibility is that structural information is represented independently from its sensory consequences, as do grid cells in spatial experiments. Such factorised representations allow components to be recombined in many more ways than were experienced, thereby enabling novel inferences. To test that, we performed Study 2.

In the Study 2, we presented eight visual objects, A, B, C, D, A', B', C', D', in a total scrambled order, they formed two true sequences (e.g., sequence 1: A->B->C->D; sequence 2: A'->B'->C'->D'), but all of the true transitions (e.g., C->D; A->B; B->C) were disrupted in the order of visual experience, so that correct sequences (e.g., A->B->C->D) could only be inferred using structural knowledge (**Figure 1a**). Here, the structural knowledge was acquired through learning a mapping between presentation order and structural position of objects in Day 1 training, for example, the first thing presented in the visual flow, D'->B->C'->C, that is D', should be put in the position 4 of sequence 2. Notably this training happened the day before scanning and with different stimuli. Subjects were therefore trained on the structure of the task but not the stimuli. This mapping pertained across Day 1 training and the Day 2 MEG experiment such that the structural knowledge was shared. Participants were extensively trained about this structural rule on Day 1. Day 2 took place in the MEG scanner, but now with novel objects. Participants were first put into a 5 minutes rest period before any visual exposure to the objects on Day 2. Participants then performed a functional localizer task. Next, stimuli were presented in the jumbled order described above, same as Day 1 training. This was followed by another 5-minute resting period. Notably, to test for neural representation of structure knowledge, participants were shown the stimuli again in the end, but now asked to perform one of two judgments as each stimulus appeared.

On *position judgment* trials, they were asked to indicate the position (i.e., position 1, position 2, position 3 or position 4) of the stimulus within the sequence it belonged to. On *sequence judgment* trials, they were asked to indicate which sequence (i.e., sequence 1 or sequence 2) the stimulus belonged to.

We wanted to know whether replay could infer a new sequence that it had never experienced. We first trained a set of classifiers to recognize individual objects using data from functional localizer - "stimulus code". Using these stimulus code classifiers, we examined the data from the resting period following the learning phase. Same as Study 1, we found evidence for forward sequenceness following a rule-defined order (**Figure 1b**) but not the order of visual experience (**Figure 1c**). Notably, this inferred stimulus order cannot emerge from simple associative mechanisms as no correct associations are present in the visual sequence. Instead this reordering must rely on a transfer of prior structural knowledge.

How is the structural representation related to sensory information during rest? To address this, we trained two additional sets of classifiers. One was trained to recognize the position of each stimulus within its respective sequence based on *position judgment* data - "position code". The other was trained to recognize which sequence each stimulus belonged to, based on *sequence judgment* data - "sequence code". As an extra precaution against contamination of position and sequence codes by coincidental common sensory features, we regressed out the corresponding stimulus code from each position and sequence classifier. We then applied the three sets of decoders (stimulus, position and sequence codes) to resting period following learing. We found both the position and sequence codes were systematically activated 40-60 ms before the corresponding stimulus code (**Figure 1d, e**), and position and sequence codes were co-activated during rest ($p < 0.0001$). These results are consistent with a model shown in **Figure 1f**, where each individual object representation in a replay event is preceded by both sequence and position representations. We speculate that these abstract structural representations contribute to retrieving the correct object for the current place in the sequence.

Is structure knowledge explicitly and independently represented before task? In rodents, prior to experience with a novel environment, hippocampal place cells play out spontaneous trajectories, which later map onto real spatial trajectories when the environment is experienced ('preplay'). Preplay may encode general abstract information about the structure of space. Our task allowed us to ask whether previously learnt abstract representations of task structure is replayed before task experience (like preplay). We performed the same sequenceness analysis to the first resting period at the beginning of the MEG scan, using transition matrix pos1->pos2->pos3->pos4. We found significant reverse sequenceness, peaked at a 30 ms time lag (**Figure 1g**), suggesting structure information is independently and explicitly represented before task.

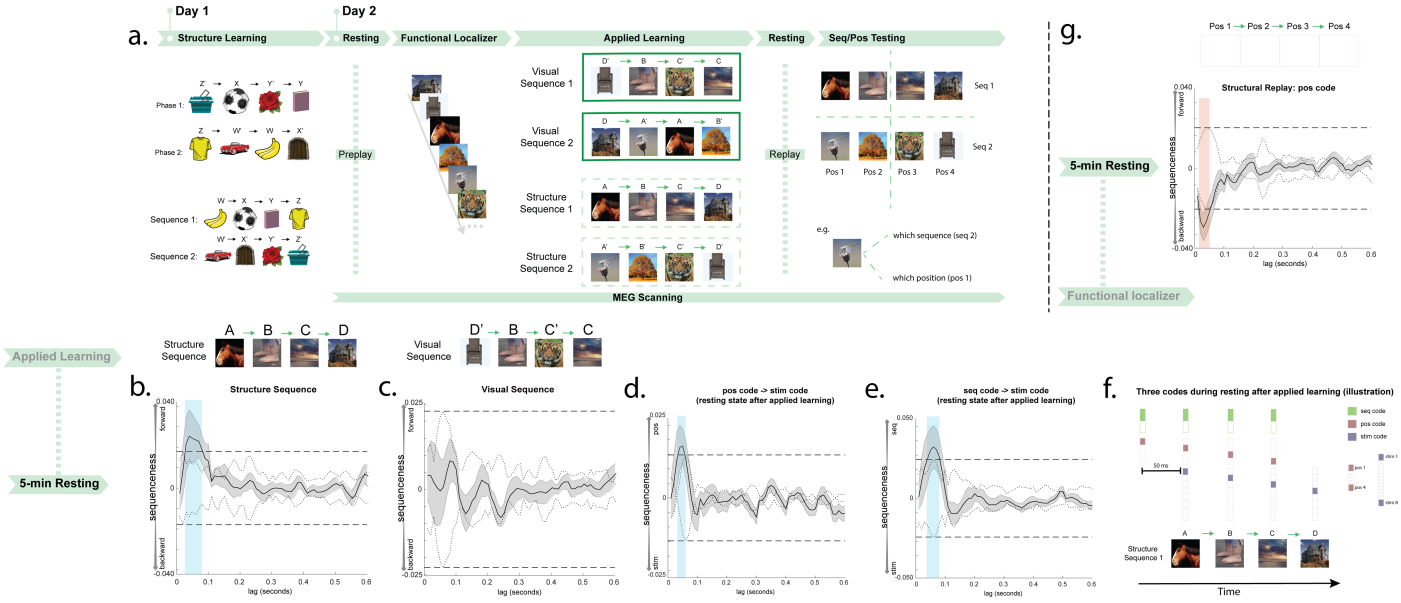

Figure 1. Task design & Replay generalize structure to new experience with factorized representation. **a,** Task design of Study 2. **b,** There was evidence for replay of the task-defined sequence, blue colour indicates significant time lags. **c,** But no evidence for replay of the visually experienced sequence. Reactivation of *pos* code (**d**) and *seq* code (**e**) consistently preceded in time of its corresponding *stim* code with a 50 ms lag. **f,** Illustration of how structure code guides replay of its corresponding sensory code into right order. **g,** Structural replay during the rest period before task experience, like preplay, red colour indicates significant time lags.