

ORIGINAL ARTICLE

Transformation of Event Representations along Middle Temporal Gyrus

Anna Leshinskaya¹ and Sharon L. Thompson-Schill

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA

Address correspondence to Anna Leshinskaya, Department of Psychology, University of Pennsylvania, 425 S. University Ave, Stephen A. Levin Bldg., Philadelphia, PA, 19104, USA. Email: anna.leshinskaya@gmail.com

Abstract

When learning about events through visual experience, one must not only identify which events are visually similar but also retrieve those events' associates—which may be visually dissimilar—and recognize when different events have similar predictive relations. How are these demands balanced? To address this question, we taught participants the predictive structures among four events, which appeared in four different sequences, each cued by a distinct object. In each, one event (“cause”) was predictably followed by another (“effect”). Sequences in the same relational category had similar predictive structure, while across categories, the effect and cause events were reversed. Using functional magnetic resonance imaging data, we measured “associative coding,” indicated by correlated responses between effect and cause events; “perceptual coding,” indicated by correlated responses to visually similar events; and “relational category coding,” indicated by correlated responses to sequences in the same relational category. All three models characterized responses within the right middle temporal gyrus (MTG), but in different ways: Perceptual and associative coding diverged along the posterior to anterior axis, while relational categories emerged anteriorly in tandem with associative coding. Thus, along the posterior–anterior axis of MTG, the representation of the visual attributes of events is transformed to a representation of both specific and generalizable relational attributes.

Key words: associative learning, events, long-term memory, middle temporal gyrus, predictive learning, relational categories, visual statistical learning

Introduction

The typical predictive relations among events form an essential component of world knowledge, even with respect to everyday objects. To understand what it is for something to be a “light switch,” one must understand the effect of flipping it; to be a “poison,” the effect of ingesting it; and to be a “plant,” the importance of watering it. It is the way these objects participate in contingencies per se that is essential, because to say that plants often get watered, or sometimes wilt, does not capture the essentially contingent property that the plant will wilt if it is not watered (Gentner 1983; Pinker 1989; Mumford 1998; Jones and Love 2007; Goldwater and Gentner 2015). Here, we investigate the cognitive and neural mechanisms supporting long-term memory of predictive relations, a relatively neglected

topic in human cognitive neuroscience (cf., Schapiro et al. 2012; Garvert et al. 2017).

There are two challenges that such representations present to the cognitive system. First, predictive relations can hold among events which look nothing alike—such as flipping a light switch and a lamp turning on (Hindy et al. 2016; Kok and Turk-Browne 2018). Events that look unlike must sometimes become related, while events that look alike must sometimes be dissociated (such as two switches which turn on different lamps). Thus, associative representations pose a distinct, often conflicting, demand from representing events' visual properties. Both are important to cognition generally: One would want to relate switch flipping and lamp lighting while still recognizing the visual similarity between two lamps never seen together. As we review below, both functions have been attributed

to the ventral temporal lobes in separate studies, creating a puzzle regarding whether both of these functions can be simultaneously accomplished in the same neural area. We specifically test this idea here.

The second challenge is that representations of predictive relations must be generalizable to be a useful component of semantic memory. Objects that participate in similar relations—such as two different switches that, when flipped, both control lamps—can be seen as relationally similar, even if those two switches are not themselves associated. On the other hand, switches that turn on lamps versus turn on ceiling fans can be seen as distinct. Likewise, one can identify the functional similarity between different coffee makers, telephones, and umbrellas. This is distinct from the first challenge of representing the individual predictive relations because it requires one to see the similarity among multiple individual relations, which are not themselves predictive of each other: One might never see the two switches, coffee makers, or telephones in the same place. However, they can be recognized as conceptually similar in terms of the structure of event relations they take part in. Thus, this requires recognizing not only the specific relation between two events but also recognizing, by analogy, multiple event relations as similar or different, an ability termed relational categorization (Goldstone et al. 1991; Markman and Gentner 1993; Gopnik and Meltzoff 1997; Markman and Stilwell 2001; Jones and Love 2007; Christie and Gentner 2010; Kemp et al. 2010; Stuhlmüller et al. 2010; Corral and Jones 2014). Our understanding of the neural mechanisms of relational categorization is extremely limited, as is our understanding of how we build such categories from event experience. Here, we test the idea that specific associative knowledge (light switch flipping and lamp turning on) is related to, and therefore recruits similar neural systems as, relational categorization (relating two switches that both control the same lamp).

We test these ideas by measuring all three kinds of representations simultaneously: visual similarity (“perceptual coding”), specific predictive relations between pairs of events (“associative coding”), and relational categories of similar specific relations (“relational category coding”) by fitting different kinds of similarity models to neural responses as participants view events and recall their predictive structure. We hypothesize that cognitively, relational categories could be built by relying on specific predictive representations, and thus, we anticipate a close relationship between neural representations of them. On the other hand, we expect perceptual coding to diverge from both associative coding and relational categories, as these functions are at odds.

Prior work offers an elegant way to probe long-term memory of specific predictive relations by examining which areas show correlated responses to individual presentations of visual stimuli after learning their association (e.g., Sakai and Miyashita 1991). Only after learning, visually responsive neurons previously tuned specifically to stimulus A increase their response to associated stimulus B, even when A and B are no longer presented together. This signature captures an important part of what it means to represent a specific relation: Given that perceptual similarities between associated stimuli are controlled, the only reliable commonality between associated pairs is the fact of their association. Thus, a common response between them would seem to represent this fact. We use the term associative coding to designate this signature.

Neurophysiological research has found associative coding signatures in various higher-level ventral visual stream areas,

specifically anterior–medial aspects of macaque inferior temporal (IT) cortex (Miyashita 1988; Sakai and Miyashita 1991; Higuchi and Miyashita 1996; Erickson and Desimone 1999; Messinger et al. 2001; Naya et al. 2003). These areas span macaque area TE, an apex of the ventral visual stream, and perirhinal and entorhinal cortices, more associated with memory. In partial accordance with this work, human functional magnetic resonance imaging (fMRI) has found evidence of associative coding in ventral stream areas like the parahippocampal place area (PPA) and fusiform face area (FFA) (Polyn et al. 2005; Turk-Browne et al. 2010; Zeithamova et al. 2012; Favila et al. 2016; Senoussi et al. 2016) but also in earlier visual areas such as V1 (Hindy et al. 2016)¹.

As we noted earlier, it is a puzzle how associative coding could take place in areas responsible for representing visual features. How could ventral stream areas simultaneously distinguish faces from houses (e.g., Haxby et al. 2001) and represent an associated face and house similarly? Surely, these functions would interfere with each other. On this basis, we predict that associative representations are more strongly represented outside of the specific areas that encode their visual features (i.e., perceptual coding), even if still within the temporal lobe. However, past work has not directly compared these functions, so it remains unknown whether associative coding signatures indeed are found primarily or most strongly in the same neural areas that represent visual features.

Nonetheless, in-line with this intuition, associative coding has also been found outside of ventral stream areas, notably in the hippocampus, to a sometimes stronger or fuller degree (Schapiro et al. 2012; Hindy et al. 2016; Kok and Turk-Browne 2018). For example, Hindy et al. (2016) found that hippocampal representations capture more of the full sequence of a set of events than visual prediction in V1, and Kok and Turk-Browne (2018) found that V1 responses are dominated by an on-screen stimulus more than what is predicted from it. However, these findings regarding V1 do not address the rest of the ventral stream. Furthermore, hippocampal responses may be limited to recently learned, preconsolidated knowledge. Others have reported that ventral stream areas broadly defined are not the strongest ones to represent predictive content and find different cortical areas that do (Kuhl and Chun 2014; Long et al. 2016). Overall, both sets of findings bolster our prediction that associative coding and perceptual coding diverge neurally. We test this idea directly by creating orthogonal models of these forms of coding and directly comparing their signatures.

There is much less prior work regarding the neural representations of relational categories, where multiple specific predictive relations may be seen as similar or different to each other. Only two studies, to our knowledge, have investigated relational categories in the brain. Frankland and Greene (2015) probed agent (i.e., action instigator) versus patient (i.e., action recipient) roles as expressed syntactically in language; they examined where sentences like “the truck hit the ball” elicited similar neural responses to sentences like “the ball was hit by the truck” (same relation), but different from “truck was hit by the ball” (different relation but similar surface features). They found this pattern specifically in the left lateral superior temporal

1 Some of this work adopts the analytic approach of seeing correlated multivoxel responses between cue and outcome stimuli, while others use classifiers trained on outcomes to test neural patterns in response to cues. We treat these as equivalent signatures.

cortex (near superior temporal gyrus), along with information representing which object participated in which role. Using a somewhat similar paradigm, Wang et al. 2016 found a nearby area, among others, but did not do inferential testing of cortical location.

There could be important differences between relational representations arising from syntactic analysis as opposed to retrieval from long-term memory. In this work, the ball and the truck served agent and patient roles equally often, and so the role to which each concept was assigned had to be determined through the syntactic evidence on that particular trial. Here, we were interested in the semantic memory of typical roles, such as an object that is always an agent (e.g., always hits the ball) or always a patient (is always hit by the ball) as retrieved from memory. However, we tested the region identified by Frankland and Greene (2015) to see whether it might be involved in both functions.

To accomplish these research aims, we taught participants various pairwise predictive relations among events, presented in four distinct sequences, where each distinct sequence was cued by a different continually present object (Fig. 1). These four distinct sequences each involved a similar set of four events, but the relations among the events could vary. Each sequence contained one strongly predictive event pair, which we call the “cause” and the “effect” (rather than a cue and an outcome, as there were no prespecified “cues” in our paradigm)². For example, as shown in Figure 1, in the sequence with the blue object, the object tilting (cause) reliably preceded the light flash appearing (effect). Between sequences and objects, we varied which particular events served as the cause and the effect, so as to create relational categories among them. In one category, the objects were “causers”: Both the blue object and the yellow object had movements as the cause and the light flash (an ambient event) as the effect. In the other category, the objects were “reactors”: The light flash was now the cause, and the objects moved in response to it (green and red objects in Fig. 1, which tilt or move a detachable part following the light flash). The two objects in the same category always exhibited different movements (whole-body tilting vs. moving a detachable part) to ensure that surface similarity went against the grain of the relational categories. Participants were then scanned about a week following learning, to ensure we probed consolidated long-term memories. During the scan, we did not show any sequence information but, rather, had participants retrieve it from memory as they viewed the individual events in random order alongside the objects (Fig. 2).

We defined three similarity models that specified which pairs of conditions should elicit more versus less correlated neural response patterns, which we then used to fit to neural data across participants (Fig. 3 and Supplementary Fig. S3). To probe associative coding, we measured the extent to which the individual cause and effect events elicited correlated neural patterns, relative to weakly predictive pairs, within the context of the same object. To probe relational categories, we compared events in the context of different objects to each other, testing whether there is a broad similarity between seeing events with object A (whose movements are followed by light flash) and object B (which moves a detachable part followed by light flash),

relative to object C (which tilts after a light flash). This kind of representation thus captures a generalized relation: the relation that a light flash is predictable, even across visually different causes (part-move vs. tilting) and across distinct contexts, but different from a light flash being the predictor of movement events (tilt and part-move). Finally, to probe perceptual coding, we measured the correlation of neural responses to the same event versus different events as seen in the context of different objects—for example, whether a light flash surrounding object A was similar to a light flash surrounding object B, relative to a movement of object B. We measured how well each of these similarity models accounted for the neural similarity structure in various parts of the brain and tested how the cortical locations of these three forms of representation were related to each other.

Methods

Participants

Participants were recruited from the University of Pennsylvania community via the Experiments@Penn website. Procedures were approved by the Institutional Review Board at the University of Pennsylvania, and all participants provided written informed consent. Participants in Session 1 were required to be between 18 and 35 years old, with no history of neurological disorder, and right-handed. To continue to Session 2, they had to be eligible for MRI following detailed screening and achieve > 80% performance on a forced-choice test at the end of the Session 1 learning task. This was done to keep constant the amount of training exposure while maintaining near-ceiling accuracy (see Results). Two participants served as a pilot sample for parameter testing and were not included in reported analyses. Ninety-seven additional participants performed Session 1. Of these, 9 did not meet MRI eligibility criteria during the detailed screening; 1 was excluded due to participating in a related prior experiment; 10 were unable to be scheduled for Session 2 within the targeted time-window; and 24 were excluded based on their performance on the training task. This exclusion rate is relatively high but follows the performance cutoff specified in the preregistration (see Procedure). Fifty-three participants took part in Session 2 and underwent fMRI. Of these, a total of 17 was excluded due to the following reasons: substantial misunderstanding of the task (1), failure to form relational categories as assessed on a postscan measure (3), technical glitch causing data loss (1), completion of fewer than 8 runs due to delays or discomfort (6), and excessive motion (6). The final sample included 36 participants (25 female), with a mean age of 23 (range 18–35).

Registration

The methods of this experiment were preregistered at <https://osf.io/3mj4v/>. Major deviations from the registration are noted in the manuscript, and minor ones on a document available at this URL. Most notably, we increased our sample size from 24 to 36 following two major unexpected outcomes: 1) inability to find associative coding in the medial temporal lobe (MTL) as expected based on prior literature and 2) inability to find any region at the whole-brain-corrected level representing relational categories. The latter prevented us from being able to test our hypothesis about where such representations would be localized. However, following the addition of 12 participants, these facts of our data did not change; thus, the sample size increase

2 In fact, we find here and elsewhere that participants see strongly predictive events like these as causally related (Leshinskaya and Thompson-Schill, 2019). However, this terminology is not specifically necessary except for convenience.

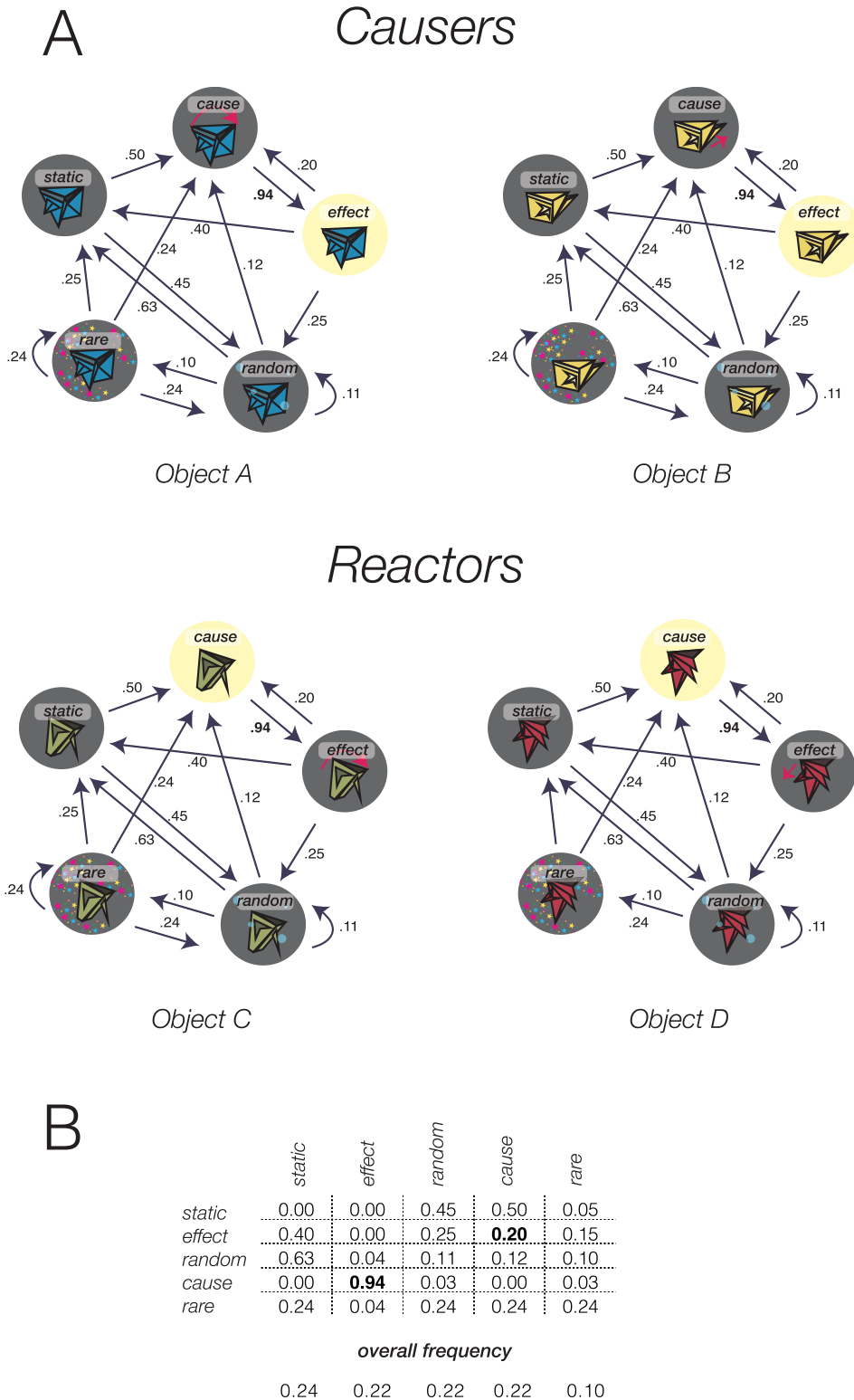


Figure 1. (A) Transition probability structures (Markov chains) used to govern the appearance of any particular event following any other during the sequence presentation in the training task. Weights (numbers on the arrows) specify the probabilities with which an event follows another. The transition relations among the abstract roles were the same for all sequences, but the particular event assigned to the cause and effect roles differed between the causers and the reactors. The identity of the ambient event serving as cause/effect was counterbalanced across participants. (B) The transition probability structure as above, shown in matrix form; below, overall frequency of each event. This figure is available for viewing in full resolution at <https://osf.io/v6pum/>.

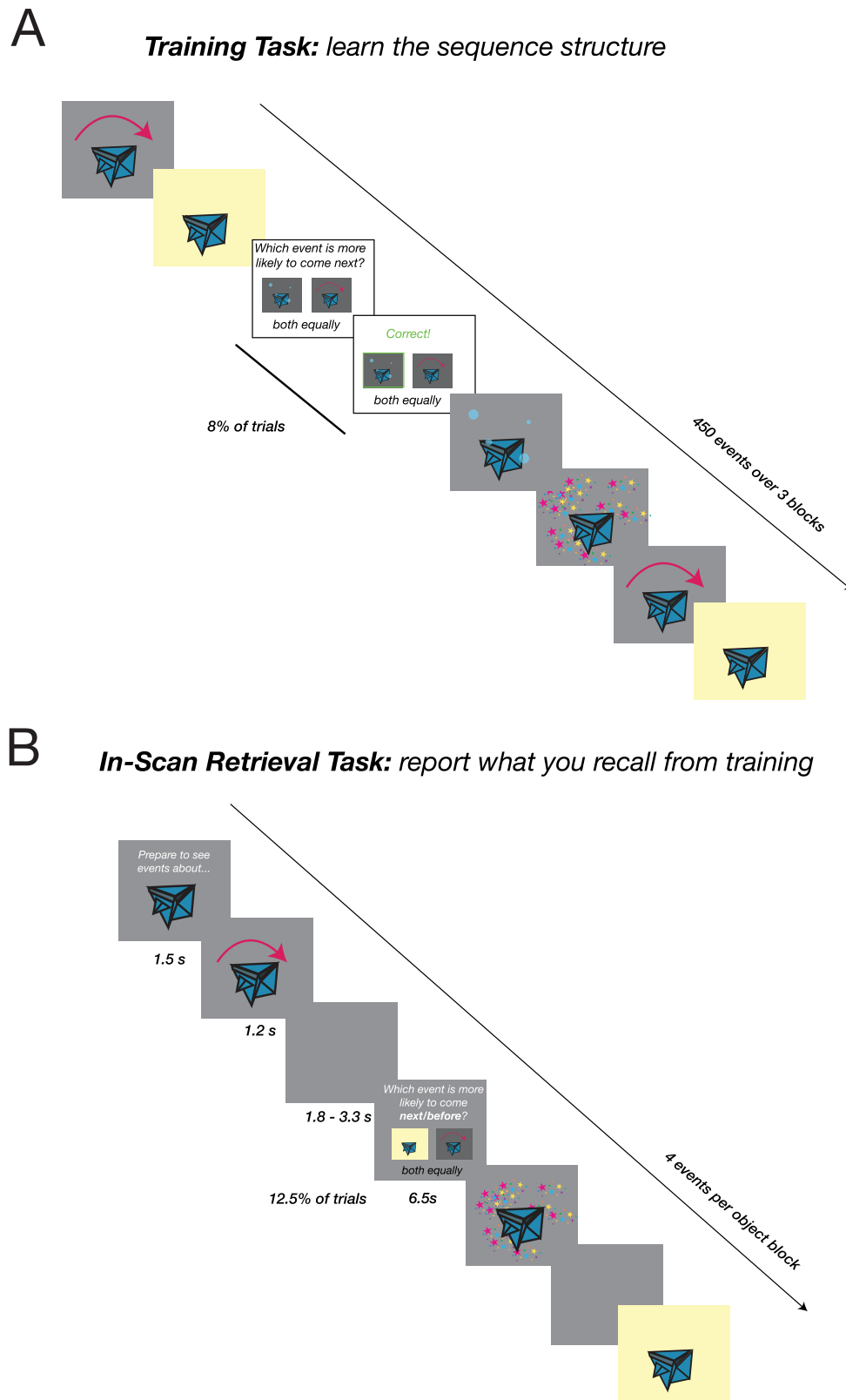


Figure 2. (A) Illustration of the training task; events are shown in sequential order (as governed by the transition probability structure shown in Fig. 1) over several minutes, interspersed with questions probing what will come next. Feedback and correct response is provided after each trial. (B) Illustration of the in-scan retrieval task. Events are shown in randomized order with delays between them, blocked by object; participants are told to recall their learning and to prepare to respond to probe questions (shown on 12.5% of trials). Half of the probe questions ask what would typically come next, the other half what would typically come before, so that specific response preparation is not possible. No feedback is given. This figure is available for viewing in full resolution at <https://osf.io/czwaha/>.

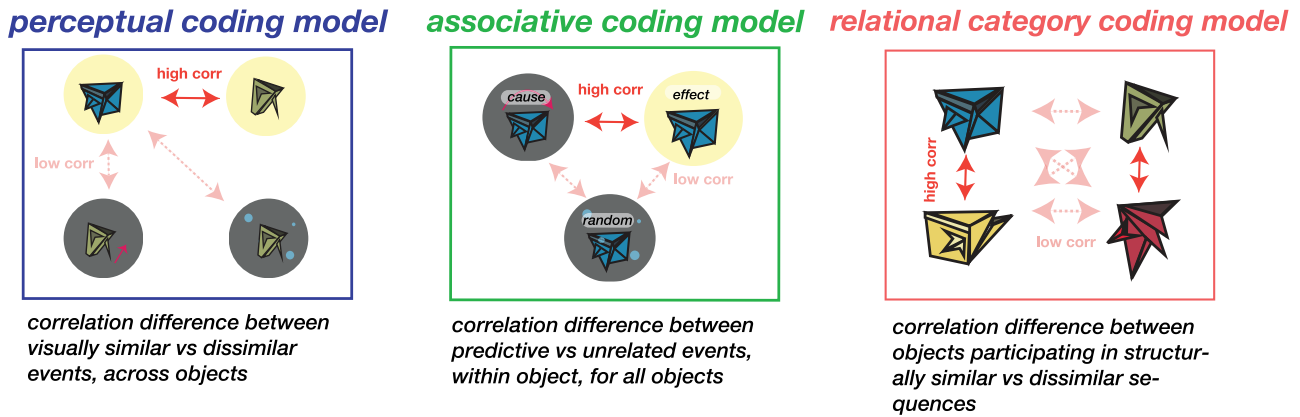


Figure 3. The three kinds of representations tested with MVPA. Each kind of representation is characterized by a similarity model. Voxelwise correlations were computed between relevant pairs of conditions; we then subtracted pairs expected to be less correlated (labeled “low corr”) minus those expected to be highly correlated (labeled “high corr”), according to each respective model. Thus, all models were translated into a correlation difference value. The signature of perceptual coding is that stimuli which are visually similar elicit more correlated neural patterns relative to those which are visually dissimilar (using correlations among pairs of events that did not appear in the same sequence). This was performed irrespective of the role of that event in the sequence (see Methods for more details). The signature of associative coding is that events that reliably predicted each other during training elicit more correlated responses than those that occurred in the context of the same object but did not reliably follow each other. The high-correlation pair was always the cause and effect event within an object, and the low-correlation pair were always the cause and random and effect and random. This was repeated for each object. The signature of relational category coding is that objects cueing structurally similar sequences (e.g., the two causers) elicit correlated neural activity, while those cueing dissimilar sequences (e.g., a causer and a reactor) elicit less correlated activity. This figure is available for viewing in full resolution at <https://osf.io/ahe5g/>.

is not likely to have inflated the significance level of analyses we do report. Inclusion criteria were as prespecified, but we additionally required that participants showed evidence of having formed relational categories, as this would otherwise hamper our ability to find such representations in cortex. Deviations and follow-up analyses that were not preplanned are indicated in the Methods and Results sections.

Overview of Session Structure

Participants completed two sessions, which were 3–11 days apart ($M = 6$)³. Session 1 took about 2 h and involved a training task (see Procedure). At Session 2, participants reviewed what they had learned in Session 1 by repeating a shortened version of the training task, then underwent fMRI scanning while performing a retrieval task, and answered a postscan questionnaire.

Stimuli

Stimuli are illustrated in Figure 1. They consisted of four novel geometrical objects, each embedded in five distinct types of animated events, presented as GIFs: bubbles, stars, light flashes, movement (either whole-body tilt or local part movement), and static (object still on-screen). Each event comprised 12 100 ms frames (total duration 1200 ms), except the static event (total duration 2400 ms). Frames were hand-drawn using Adobe Illustrator and concatenated into GIF files using MATLAB (MathWorks).

For the training task, these event stimuli were concatenated into 450-event-long sequences (in which only one object was presented); this created the “object contexts.” The order of events in each sequence followed a specific structure, as summarized in the pairwise transition matrix shown in

Figure 1B (and in graphic form in Fig. 1A). This matrix specifies the conditional probability of moving into any specific state at any trial n given the state at trial $n-1$. Sequences for each participant were generated probabilistically using a weighted walk, where the probabilities of adding events to the sequence were specified by this transition matrix. We ensured that the generated sequences closely matched this specified probability structure by checking that, in each generated sequence, the average absolute difference in all pairwise transition probabilities was below 0.00004 and the standard deviation (SD) of state frequencies of the cause, effect, and random were below 4. The actual average obtained transition matrix was nearly identical to the specified sequence.

Although the same transition matrix governed the abstract structure of the predictive relations in all four sequences, the way that the events were assigned to this structure varied (Fig. 1A). In all cases, a strong predictive relation was held between two events, the cause and the effect, such that the cause is followed by the effect with a 94% probability. For two objects, the causers, the cause was the object’s movement (tilt for object A and part-move for object B), and the effect was one of the three ambient events (bubbles, stars, or light), selected for each participant in a counterbalanced fashion, but always the same for the two objects (e.g., light flash in the example in Fig. 1C,D). For the other two objects, the effect and cause events were swapped: Object C tilted following the ambient event (e.g., light flash), while object D moved a detachable part following the same ambient event (also light flash in this example). In this way, object contexts belonged to one of two relational categories, causers and reactors.

Two other events served as “random” and “rare” events, which were almost never predicted by the cause and almost never predicted the effect; the identity of these events was the same across categories. The random event was matched in frequency to the effect and cause, while the rare event was half as frequent; overall frequencies are displayed in Figure 1B. This difference in frequencies was introduced to enable comparison

³ The preregistration had indicated the window would be 9 days, but we had one exception in which a participant was scheduled with a delay of 11 days.

to a planned follow-up, in which categories were based on frequency rather than contingency, but is not a manipulation of interest here. The rare events were largely excluded from analyses, as described below; comparisons largely involved the cause, effect, and random events. The static event served only to facilitate learning by providing a break in the sequence structure; it was for this reason that it was longer and slightly more frequent. This event simply showed the object still on the screen. It was not the target of any test questions, nor was it a condition in the fMRI session.

The assignment of object shapes to relational category was counterbalanced across participants, creating six counterbalancing conditions for object shape (i.e., all possible assignments to two categories). Relational category was orthogonal to object movement, as the two members of each category always had different movements. The assignment of the three event ambient types (light flash, bubbles, and stars) to be the “effect/cause,” random, and rare was also counterbalanced across participants, creating six other counterbalancing conditions, which were paired randomly with the shape counterbalancing conditions.

During fMRI, participants saw individual event GIFs for each of the four nonstatic events (cause, effect, random, and rare) in the context of each of the four objects, creating 16 conditions. As described below, this presentation was in random order, rather than following the transition probabilities as during training.

Procedure

Session 1

Participants were introduced to the four object shapes and told that their task was to learn which events are likely to follow which others, in the presence of each object. Thus, it was very explicit to them that they were expected to learn the transition structure governing the sequences and how these structures might vary by context and that they would be tested on their knowledge throughout. Session 1 took place over 1.5–2 h; thus, participants had extensive exposure to the displays.

The 450-event sequence for each object was split into a preview block and three task blocks. The preview showed the first 50 events (~1 min) from each object's sequence, with the order of sequences/objects random. The following three blocks showed the remaining 400 events from each sequence. These were presented in sets by block number, such that participants saw block 1 for all four objects, then block 2 for all four, and so on, with the order of sequences/objects randomized uniquely at each block.

In these task blocks, the videos were interspersed with intermittent questions, which probed the participant to decide what event is most likely to come next (Fig. 2A). Thus, participants' learning was probed explicitly throughout exposure. The response options showed static images of two other events (not itself or rare⁴) and a “both equally” option. For example, following the presentation of a cause event, participants would choose between the effect, random, or both equally. The both equally option was correct for events with close or

equal transition probabilities (within 5%), but otherwise the correct option was defined as whichever event had the higher conditional probability. For example, following the cause, the correct answer was always the effect. However, following the effect, cause and random were equally likely; and following random, cause was more likely. Presentation side of the two event options was randomized, with both equally appearing below. Participants received feedback following their response, showing them which response was correct if they were incorrect. For each object, there were 10 questions pertaining to the cause, effect, and random events and 6 pertaining to rare, creating 36 questions total, distributed randomly over the 400 events. To create the three blocks, the sequences were split so that each one contained 12 questions for each object (hence these segments could vary in the number of events).

Following each block, participants saw forced-choice tests probing their knowledge about each object's sequence; thus, they saw these tests three times. Each question showed two videos side by side, where each video contained a sequence of two events (e.g., tilt followed by light flash vs. tilt followed by bubbles). Participants had to choose which of the two videos was most typical. There were seven trials per object, including three trials comparing cause–effect to effect–cause, one trial comparing cause–effect to random–effect, one trial for cause–effect to random–cause, and two filler trials (effect–cause vs. random–effect and effect–cause vs. random–cause), which balanced the number of times the cause–effect pair and the effect–cause pair were shown overall. Overall accuracy was shown after the test on each block. Participants had to obtain 80% or higher on the nonfiller questions by block 3 to continue to Session 2. The emphasis on cause–effect order knowledge was because of their importance for the relational categories.

Session 2

Participants reviewed what they learned with a similar training task as in Session 1, but shorter: A total of 18 questions was shown per object over 240 events, split over two blocks. During fMRI scanning, participants performed a retrieval task in which they saw individual events separated by blank screens and in randomized order except blocked by object. They were intermittently asked questions probing their memory of typical event order from training. We expected them to retrieve associated events during viewing in order to be prepared for these questions. We anticipated this would increase our ability to detect associative memory representations. It was thoroughly emphasized to participants that the order of events during this task was purely random and no longer informative.

Figure 2B illustrates the retrieval task. Each block began with a cue (1.5 s) showing a still image of the object, followed by four of that object's events (these could be any of the four events in any order, but with no more than two of the same event in a row). Each event (1.2 s) was followed by a delay (blank screen) with a duration of 1.8, 2.3, 2.8, or 3.3 s. On 12.5% of trials, this delay was followed by a question (6.5 s). Half of the questions asked what is likely to come next; the other half asked what is likely to come before. Because the questions could appear any time, and were relatively fast, participants were told that it would be advantageous if they recalled their contingency knowledge as the events appeared. However, because the questions could probe before or after knowledge, participants could not prepare any specific response. Each question had three response options, as in the training task: Two options were images of two other events (except itself or rare), and one was both equally. The

4 During piloting, we found that including questions regarding the rare event increased task difficulty, given the limited evidence they saw about it, and excluding it simplified and focused the training task. It was also not clear that the ability to select effect over rare reflected knowledge that the rare event had an overall lower base rate, rather than transition knowledge specifically.

presentation side of the two event images was randomized. Unlike the training task, here no feedback was given, except an overall score at the end of each run. Order of object blocks was randomized.

Each of the 16 events (cause, effect, random, and rare for each of the four objects) was shown exactly four times in each run, once with each delay duration. These were arranged randomly into four-event blocks (with the constraint that no event could repeat more than twice in a row within a block). There were 10 runs over the entire experiment, and thus 40 repetitions of each event. Questions were distributed across the entire 10 runs, ensuring that there were 5 questions for each event (12.5%), and half were “after” and half were “before” questions. For analysis of fMRI data, question periods were modeled separately and not further analyzed.

Following scanning, participants completed a questionnaire asking whether they thought the objects could be naturally grouped into categories (yes/no) and, if so, how many and on what basis (freeform text entry). They were then shown draggable images of each object shape and asked to arrange them on the screen such that the ones they thought were most similar were closer together and the ones they thought were most different were furthest apart. Finally, they answered questions about their perceptions of causality and animacy. The animacy question asked, “To what extent did the four objects you learned about seem like animate, living objects (animals/people) versus inanimate (nonliving objects like artifacts)?” with a 1–5 response scale where the endpoints were labeled “Definitely Animate” and “Definitely Inanimate,” with the side of the scale of these labels randomized. Another question probed their perception of causality regarding the “causer” objects, “For two of the objects, their movements predicted the occurrence of another event (e.g., the appearance of a light flash, bubbles, or stars). To what extent did you perceive this relationship as causal? Did these objects seem to cause this event?” with a similar 1–5 response scale whose end points were labeled “Definitely Causal” and “Definitely Not Causal” and the ends of the scale randomized. A third question probed causality about the reactor objects (“Did the event seem to cause the object to move?”).

fMRI Acquisition Parameters

fMRI data were acquired using a Siemens Magnetom Prisma 3T scanner at the University of Pennsylvania, using a 64-channel coil. Anatomical volumes were acquired with a T1-weighted MPRAGE sequence with $0.8 \times 0.8 \times 0.8$ mm voxel resolution, 256 mm field of view, time repetition (TR) = 2.40 s, and time echo (TE) = 2.24 ms. Functional data were acquired with a multiband echo-planar imaging (EPI) blood oxygen level-dependent (BOLD) sequence using 72 interleaved slices with a multiband acceleration factor of 3, $2 \times 2 \times 2$ mm in-plane voxel resolution, 220 mm field of view, TR = 2.0 s, TE = 30 ms, and flip angle = 75° . Slices were aligned to the posterior–anterior axis of the hippocampus (following Schapiro et al. 2012).

fMRI Preprocessing

Data were preprocessed using AFNI software (Cox 1996). Slices in each volume were corrected for acquisition timing using Fourier interpolation (3dTshift). Each volume was spatially aligned to the fourth volume of the first scan to correct for motion (3dVolReg). Image intensities were normalized (scaled to range from 0 to 100), and linear and polynomial slow trends up to the

third level were removed. Data were spatially smoothed using a Gaussian kernel of 3 mm full-width half-maximum. Runs in which displacement from the first exceeded 3 mm were excluded; if more than two were excluded, the dataset was discarded for that participant (and replaced to fulfill the counterbalancing set). Included runs were then concatenated into one time series and entered into linear modeling. Anatomical volumes were spatially aligned to the first functional volume in-line with the rest of the functional data and then spatially transformed to Talairach space. Parameters for Talairach transformation were then applied to functional scans for volume analyses.

Linear Modeling

The objective of linear modeling was to estimate individual participants' neural response to each of the different event types (i.e., to determine how strongly each voxel responded to each event). The response strength at each voxel serves as the input to later similarity models, described below, which are the critical analyses. Two linear models were fit to the data. The Object model included regressors for each object (A–D), which spanned cue periods, event trials, and delay periods, but excluded question periods, which were modeled with a separate regressor (this was to avoid including contaminants such as decisions, motor responses, and irrelevant visual stimuli to estimates of response patterns). Derivatives of the six motion realignment parameters (four directions and two rotations) were also included. The Event model included regressors for each of the 16 different events (cause, effect, random, and rare for each of the four objects) and, as above, a regressor for all question periods and 6 motion realignment parameter derivatives. In both models, volumes with motion outliers (those with > 0.15 mm displacement from the previous) were excluded.

Regressors were created by convolving the time courses of each condition in each run with a gamma-shaped hemodynamic response function. The convolved time courses were then used as predictors in a least-squares linear regression over the time courses of BOLD signal in each voxel (3dDeconvolve). This produced a map of regression coefficients for each condition, and their respective t-values, reflecting the slope of the relationship between that voxel's signal and the occurrence of that condition. The t-value maps were used in all subsequent analyses.

Anatomical Surface Analysis

Anatomical volumes were converted to surface maps for surface-based searchlight analyses. Surfaces were created using the Freesurfer function recon-all (Fischl et al. 1999), which used intensity gradients to segregate white and gray matter and generate inflated cortical surface maps for each individual participant. This algorithm also performed segmentations of medial temporal lobe areas which were used in region of interest (ROI) definition (see below). Interindividual alignment of surface maps and alignment of functional data to surface maps were performed using AFNI (mapIcosahedron) and algorithms implemented in the Surfing toolbox (Oosterhof et al. 2014; Oosterhof et al. 2011).

Multivariate Analyses

Multivoxel pattern analyses (MVPAs) were performed on the outputs of linear models (t-values reflecting the strength of response in each voxel to each condition). The goal of these

analyses was to probe the fit of the three similarity models of interest (associative coding, perceptual coding, and relational category coding) by comparing which pairs of conditions elicited relatively more correlated versus less correlated responses, across certain sets of voxels. These analyses were performed in various parts of the cortex, described in more detail below. For example, for whole-brain searchlight analyses, the sets of voxels were defined as spatially contiguous neighborhoods tiling the cortical surface. However, the calculations were always the same.

The three similarity models each specify which pairs of conditions should elicit more versus less correlated multivoxel responses (Fig. 3). In all cases, pairwise correlations among the conditions of interest were computed, subtracted from each other according to the model, and averaged into an overall correlation difference value. This correlation difference is taken to reflect how well each similarity model fits the neural data. Tests against 0 were performed with one-tailed t-tests, since our interest was only in signatures exhibited by a specific direction of contrast in each of the three models (as described below). In other words, we endeavored to localize specifically these directional signatures to relate them to each other, rather than any others. For example, although at least one prior report has found that some areas exhibit decreased similarity between items with similar paired associates (Favila et al. 2016), our hypotheses only concerned associative signatures in which related items become more representationally similar.

Associative Coding

The associative coding signature is that the neural response to an event should be correlated with the neural response to its associate, but less correlated with an equally frequent, but unassociated, event. We performed this analysis following prior human fMRI work (Schapiro et al. 2012). Thus, we computed the correlation of the voxelwise responses between pairs of conditions with strong predictive relations (the cause and effect events) and those with weak predictive relations (cause and random and effect and random), within object context. These particular comparisons were chosen because the stimuli assigned to be effect and random were counterbalanced across participants, thus fully controlling for stimulus identity, and were of equal overall frequency (Fig. 1B). The correlation values among the weak pairs were subtracted from the correlation values among the strong pairs within object. This was performed for each object and averaged. This average correlation difference value reflected the extent of associative coding in a region, which we refer to as the “associative coding model fit.”

Perceptual Coding

The signature of perceptual representations was a more correlated response between pairs of events that were more visually similar relative to events that were visually dissimilar. For example, viewing a light flash in the context of object A is perceptually more similar to a light flash in the context of object B than it is to bubbles in the context of object B (Fig. 3). To perform this analysis, we considered the stimulus properties of each event irrespective of its role in the sequence structures. We used the cause, effect, and random events, but not rare events, because the latter was not matched in frequency to the others. The high correlation pairs were thus the events with the same visual identity (stars, bubbles, light flash, or movement) but across object context. The low-correlation pairs were events with a

different identity, also across object context. One example of such a correlation pair is illustrated in Figure 3. A fuller example showing all visually similar pairs used in the analysis is provided in Supplementary Figure S2. Thus, overall, the perceptual model fit was evaluated as the difference in correlation between pairs of visually similar events and pairs of visually dissimilar events (e.g., the correlation between the light flash with object A and the light flash with object B vs. the light flash with object A and movement, bubbles, or stars with object B). This average correlation difference is referred to as the “perceptual coding model fit” as it reflects how well a perceptual coding signature characterized a neural region.

This analysis deviates slightly from its preregistration. We originally planned to compare correlations across objects, grouping the two objects which tilted versus moved a detachable part; we later realized the analysis would be better matched to the associative coding analysis by analyzing individual events.

Relational Category Coding

In a generalized representation of a relation, sequences in which the light flash is an effect should be more similar to each other than sequences in which light flash is a cause. Accordingly, our four object/sequence conditions could be grouped into two classes based on similar relational structure, that is, whether the object’s movement predicted versus followed one of the ambient events (causers vs. reactors). Here we used the outputs of the Object linear model to obtain a neural response pattern to each Object condition and obtained the pairwise correlation between all pairs of conditions in terms of their voxelwise t-values. We then subtracted the correlation of all different-relation objects (e.g., object A and object D) from the correlation of same-relation objects (objects A and B and objects C and D). The average difference was computed for each subject in each region and is referred to as the “relational category model fit.” As described below, we tested all three model fits in the whole brain using a searchlight, as well as in various specific ROIs.

Anatomical ROI Definition

Anatomical ROIs tested were left and right hippocampus, entorhinal and perirhinal cortices, as well as individual hippocampal subfields (CA1, CA3, CA4, dentate gyrus, subiculum, and tail), all as extracted from the Freesurfer segmentations of each individual. We also tested a left middle superior temporal region based on coordinates reported in Frankland and Greene (2015), at (−59, −25, 6), by defining spherical ROI with 123 voxels surrounding them. MVPA analyses as described below were performed on the voxels within each ROI in each participant. As described in a modification to our preregistration, we chose to use anatomically defined MTL ROIs to make better contact with prior work, particularly as we did not see searchlight effects in these areas in initial analyses. Nonetheless, we continued to not see significant effects in MTL (see Results).

Searchlight Procedure

To test the fit of the three similarity models across the entire cortex, we defined neighborhoods of contiguous voxels tiling the brain and performed MVPAs in each neighborhood. Searchlights were defined both on volume and surface data, with the latter preferred as it defines searchlight neighborhoods respecting the curvature of individual’s cortical surfaces, a more valid measure of contiguity (Oosterhof et al. 2014, 2011). We report clusters seen

in both analyses, but performed multiple comparison correction only on surface maps due to computational time constraints. Volume searchlights were used for reporting Talairach coordinates for clusters significant in the surface analysis. Each searchlight neighborhood had a radius of 3 voxels or 6 mm and included 123 voxels. An additional follow-up analysis targeting medial temporal areas used a 3 mm radius. All pairwise correlations among conditions in their voxelwise *t*-values were computed, Fisher-corrected, and subtracted and averaged according to the respective model, yielding a single value for that neighborhood reflecting model fit. Subsequently, *t*-tests were used to compute the statistical significance of model fit at each searchlight neighborhood across the group. Surface maps were created to display the value for each neighborhood at its center coordinate.

Multiple comparison correction was performed with permutation testing at the cluster level. For each individual, 10 null maps for each linear model were created by shuffling the condition labels across trials (i.e., randomizing trial-to-condition assignments; 10 were used to create sufficient variance). Then, for each of 1000 permutations, one of these null maps was chosen per participant at random, and analyses proceeded exactly as they had for real data. The maximal cluster size obtained from group-level analyses at each iteration was used to build a distribution of maximal cluster size expected by pure chance (noise), given a precluster threshold of $P < 0.001$. The observed cluster sizes in the real data were assigned a significance value based on their likelihood in this distribution; clusters above 105 mm² were significant at $P_{\text{corr}} < 0.05$.

Functional ROI Definition

We used the significant clusters from the whole-brain, surface searchlight MVPA with the associative coding and perceptual models to define individual functional ROIs. To do so, we used the group clusters as boundaries and then selected individual surface nodes in each participant by taking the largest contiguous cluster of nonzero nodes within it. These regions were then tested for fit of the other models, all of which used independent comparisons (between different trials) from those used to define the ROIs. This enabled us to test the theoretical question of whether the same or different regions enable associative and relational coding. This ROI definition approach follows our preregistered methods for doing so. The specific regions selected were based on the significant searchlight findings.

Vector of ROI Definition

To statistically assess the spatial relationships between perceptual and associative coding results, which we found in middle temporal gyrus (MTG) at a whole-brain-corrected significance level, we took a “vector of ROIs” approach (Konkle and Caramazza 2013). To do so, we defined a linear axis along MTG, respecting its boundaries along the surface curvature. We then defined spherical ROIs along this axis, taking every 10th node and defining a sphere around it with a radius of 10 mm; this ended up creating 39 partially overlapping ROIs along the posterior to anterior axis of MTG. We assessed the fit of the perceptual, associative, and relational category models in the same way as above to assess their relationships across these ROIs. As described below, this analysis served to test follow-up questions raised by our findings, notably regarding their spatial relationships. As such, it is considered a follow-up analysis and

was not preregistered. However, it is directly aligned with the aims outlined in the preregistration, to assess the relationship among areas showing signatures for associative, perceptual, and relational category coding. The way in which these relationships among our findings would be assessed was only prespecified broadly.

Results

Learning Performance—Session 1

Accuracy on questions interspersed through the training sequences was high among the included subjects ($M = 82\%$, $SD = 13\%$). For two participants, training task responses were missing due to technical glitches. Accuracy did not differ as a function of whether the object was a causer or a reactor, $t(33) = 1.04$, $P = 0.304$. Accuracy did differ as a function of the event probed (i.e., presented prior to the question asking what follows it), with the causal event being most accurate ($M = 86\%$, $SD = 13\%$), followed by the random event ($M = 78\%$, $SD = 16\%$), the effect event ($M = 72\%$, $SD = 25\%$), and the rare event ($M = 68\%$, $SD = 22\%$). These differences were significant for all comparisons between cause and others and between random and effect (Supplementary Table S1). Such differences no doubt arose because the predictive relations from the cause were by far the strongest and clearest, while the predictive relations among the other events were weaker. It should be noted that the associative coding model predicts relatively higher correlations between cause and effect than between cause and random (or between random and effect); in terms of accuracy, however, the effect event was most different from the cause event in terms of accuracy. Furthermore, the correlation across subjects between accuracy on the cause and effect events was lower ($r = 0.60$) than between the cause and the random event ($r = 0.70$) and between the effect and the random event ($r = 0.78$). Therefore, the difficulty of the training task itself is not confounded with the neural models tested (in fact, it goes in the opposite direction).

Forced-choice tests probed the ability to retrieve each cause-effect relation associated with each object and was used as a selection criterion for scanning. Included participants were thus highly accurate on each object, reaching ceiling by the last block (object A: $M = 98\%$, $SD = 0.08\%$; object B: $M = 98\%$, $SD = 8\%$; object C: $M = 98\%$, $SD = 8\%$; object D: $M = 98\%$, $SD = 8\%$), with no difference between causers and reactors ($t_s < 1$) either at the last block or on average across all blocks.

Learning Performance—Session 2

In Session 2, participants performed a review task similar to the Session 1 training. Accuracy remained high ($M = 85\%$, $SD = 13\%$) and followed a similar pattern across event types as in Session 1. Participants again reached ceiling on the forced-choice test (object A: $M = 99\%$, $SD = 5\%$; object B: $M = 98\%$, $SD = 6\%$; object C: $M = 97\%$, $SD = 14\%$; object D: $M = 98\%$, $SD = 8\%$), with no significant difference between the objects or the object categories (causers vs. reactors). While undergoing fMRI, participants were also highly accurate ($M = 74\%$, $SD = 16\%$), although less so, perhaps because questions were presented more quickly and not in the context of actual sequence presentation, thus drawing more on relatively distant memory. Accuracies for questions about the cause ($M = 67\%$, $SD = 19\%$) were relatively lower than accuracies about the effect ($M = 81\%$, $SD = 14\%$) as well as about random ($M = 80\%$, $SD = 16\%$). However, to match the fMRI analyses

performed here, we found that the correlation between participants' accuracies was not significantly higher between cause and effect ($r = 0.64$) than between cause and random ($r = 0.52$, $z(35) = 0.73$, $P = 0.459$) or between random and effect ($r = 0.37$, $z(35) = 1.50$, $P = 0.134$).

Subsequent to fMRI data collection, participants were asked to spatially arrange images of the four objects according to their judgments of how similar they were to each other. Four participants' data were missing due to technical error. We computed the screen distance between all pairs of objects. Participants reliably placed the same-category objects (the two causers and the two reactors) closer to each other than to different-category objects ($M = 50.33$, $t(31) = 12.00$, confidence interval [CI] [58.88, 41.78], $P < 0.001$). All but one considered them to belong to two categories, rather than any other number, despite no suggestion in the experiment that they should do so. Thus, the included participants spontaneously grouped the objects into two categories according to the predictive structure of their associated sequences.

Participants reliably saw the objects as inanimate, providing a mean rating below the midpoint of three on a 1–5 animacy scale ($M = 2.31$, standard error [SE] = 0.22, $t(35) = -3.25$, $P < 0.01$), though there was a wide range on this measure (1–5), indicating that some participants did see the objects as animate. They also rated the object movements as reliably causing the ambient event for the causers ($M = 4.04$, $SE = 0.20$, $t(35) = 5.27$, $P < 0.001$) and the ambient event causing the object movement for the reactors ($M = 3.93$, $SE = 0.21$, $t(35) = 4.53$, $P < 0.001$), with no difference between the two ($t(35) = 1.00$, $P = 0.32$). It should be noted that no causal language was used at any point during the experiment prior to these ratings.

Areas Exhibiting Associative, Perceptual, and Relational Category Coding

We performed MVPAs across the entire cortical surface using a searchlight procedure. These analyses tested the extent to which any given neighborhood of voxels exhibited the signatures of associative, perceptual, and relational category coding models, which specified particular patterns of pairwise similarities in the neural response to each pair of conditions (Fig. 3 and Supplementary Fig. S3).

For perceptual coding, we searched for regions that showed correlated activity between visually similar pairs of events across objects (e.g., light flash in object A and light flash in objects B, C, and D) relative to visually dissimilar events across objects (light flash in the context of object A and movement of objects B, C, and D). We found such effects across the inferior occipitotemporal cortex, spanning medial and lateral aspects (Fig. 4B). Lateral temporal areas are known to be particularly sensitive to dynamic events, with MTG showing selective responses to inanimate motion (Beauchamp et al. 2002). These previously reported coordinates (-46 , -70 , -4 , Talairach space) are very near the peak of our perceptual coding effects in the lateral temporal cortex (47 , -71 , 3).

The associative coding model specifies higher correlations among pairs of events that were predictive during learning (cause and effect), relative to pairs of events that were not predictive of each other (cause and random and effect and random), within each the context of each object. We found such representations in a diverse set of regions across the cortex, as shown in Figure 4A; this included the right MTG and lateral prefrontal cortex, left precuneus, and medial prefrontal cortex.

Some of these resemble parts of the default mode network (Buckner et al. 2008). Notably, although both perceptual coding and associative coding models characterize responses in parts of right MTG, these areas appeared nonoverlapping. We directly follow up on this observation with a vector of ROIs analysis described in the next section.

To compare the influence of associative similarity and perceptual similarity on the evoked responses, we performed a *t*-test contrasting the relative magnitude of associative coding versus perceptual coding model fits within participants. Figure 6 shows an unthresholded map of these effects to visualize the full range of differences across the cortex (Fig. 6A) as well as a map of multiple-comparison-corrected, significant clusters only (Fig. 6B). Despite the fact that more trials were used to test the perceptual model (Supplementary Fig. S3), areas exhibiting significantly stronger perceptual coding were limited to the posterior temporal cortex. Associative coding was relatively stronger than was perceptual coding in several areas, including right MTG, but also parts of the lateral and medial prefrontal cortex and medial parietal cortex.

We did not find patterns predicted by the associative coding model in medial temporal or IT regions. Volume-based searchlights broadly confirmed these findings (Supplementary Fig. S1); Talairach coordinates are reported in Supplementary Table S2. In anatomically defined medial temporal ROIs (hippocampal, parahippocampal, perirhinal, and entorhinal cortices), we found no significant effects (all $t_s < 1$, except in the right parahippocampal gyrus, in which $M = 0.013$, $t(35) = 1.32$, $P = 0.098$). We additionally tried a smaller searchlight radius but again found no effects at $P < 0.01$ uncorrected anywhere in the medial temporal lobes apart from a small cluster in left parahippocampal gyrus. Constraining searchlights to within individual subjects' anatomical MTL ROIs confirmed this result, revealing only a small cluster of 14 voxels in in left parahippocampal gyrus at $P < 0.001$ at the group level. Thus, overall, we found little evidence of associative coding in MTL, but robust evidence in other areas. As we raise in the Discussion section, the major difference from prior work is the long delay between training and testing, and we suspect that such consolidated associative representations as tested here may have a systematically different neural locus. We thus do not see this result as contradicting past findings but, rather, revealing a potentially important shift in their neural locus following long intervals between encoding and retrieval.

We did not find any region at the whole-brain level that reliably exhibited relational category coding, i.e., greater correlation when viewing objects that cued sequences with similar statistics (the same effect and cause events), relative to objects with different statistics. We also failed to find effects in an ROI centered on coordinates reported in Frankland and Greene (2015), $M = 0.004$, $t(35) = 1.07$, $P = 0.147$. This may be due to lack of power or to the specificity of these prior effects to language stimuli or their extraction from syntax.

Posterior to Anterior Functional Divisions in MTG

To directly test the spatial divergence between associative and perceptual coding, we performed the same multivoxel analyses in a vector of ROIs defined in individual participants (Konkle and Caramazza 2013). We defined this series of partially overlapping ROIs along the posterior–anterior axis of right MTG, so that we could directly test that the spatial relationship between associative and perceptual coding is statistically reliable along this

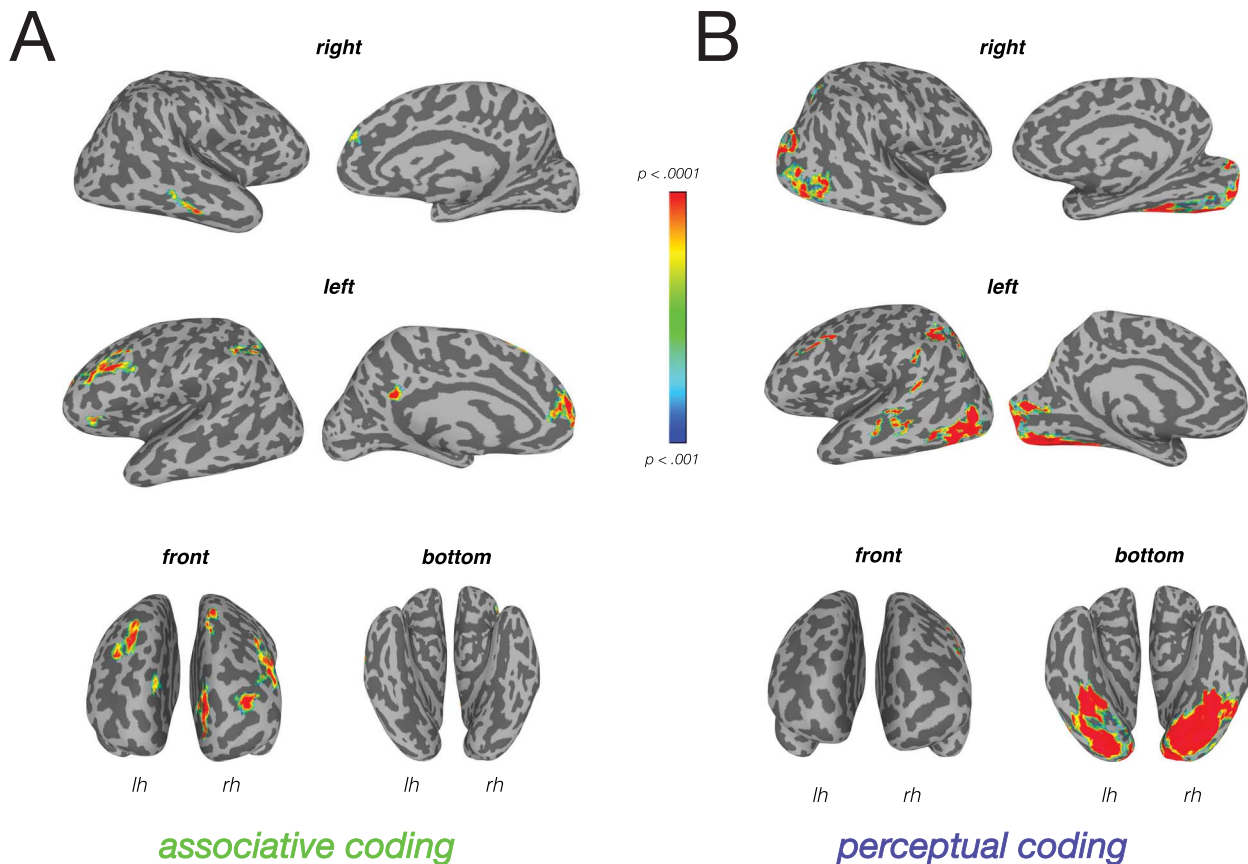


Figure 4. (A) Surface-based searchlight results for associative coding, with a voxelwise threshold of $P < 0.001$ and a cluster-corrected threshold of $P < 0.05$ (104 mm^2). (B) Surface-based searchlight results for perceptual coding, with the same thresholds as (A). This figure is available for viewing in full resolution at <https://osf.io/u8vbt/>.

posterior–anterior axis. We also used these ROIs to assess the relationship between the associative and perceptual model fits to those of the relational category model, as this was an area where two of the hypothesis-relevant models were clearly important (the perceptual and associative coding models). Because these analyses follow from our earlier observations in this experiment, they were not part of the preregistration, although the model signatures we use remain constant.

Figure 5A shows these ROIs, and the fit of each model (associative, perceptual, and relational category, in terms of correlation difference) in each ROI, arranged from posterior to anterior. This analysis revealed that perceptual coding declines, whereas associative and relational category coding increases, along this axis. This was statistically reliable: The location of the peak ROI for the associative model ($M = 19.81$, $SE = 1.53$, $CI [16.7386, 22.8725]$) was reliably anterior to the peak of the perceptual model ($M = 12.44$, $SE = 1.69$, $CI [9.0586, 15.8303]$) when compared in individual subjects ($t(35) = -3.51$, $P = 0.001$, $d = -0.76$). Fitting linear slopes to individual data along the ROIs confirmed that associative coding exhibited an overall positive slope across this axis ($M = 0.001$, $SE = 0.0005$, $CI [0.0000, 0.0021]$, $t(35) = 2.06$, $P = 0.047$) while perceptual coding exhibited a negative slope ($M = -0.001$, $SE = 0.0003$, $CI [-0.0017, -0.0005]$, $t(35) = -3.85$, $P < 0.001$), and that these were significantly different from each other ($t(35) = 3.75$, $P < 0.001$, $d = 0.87$). These analyses confirm that a divergence between perceptual and associative coding along MTG holds reliably when comparing

their locations in individual participants (something which is not guaranteed by the results of the searchlights).

Considering individual ROIs, we also found that in ROIs 21, 22, and 24, the associative model fit was stronger than was the perceptual model fit, correcting for the 39 ROIs tested ($M = 0.087$, $t(35) = 3.63$, $P < 0.001$; $M = 0.092$, $t(35) = 4.00$, $P < 0.001$; $M = 0.071$, $t(35) = 3.92$, $P < 0.001$), and that perceptual and relational category model fits did not differ from one another in any ROI (all $P > 0.05$).

We additionally explored how associative and perceptual coding relates to relational category coding across these ROIs. It is critical to test relational category representations specifically in these areas, which already show evidence of associative and perceptual coding, as it directly pertains to our central question regarding the relationship among the three similarity models, particularly in the temporal lobe (given work reviewed above).

As evident in Figure 5A, the fit of the relational category model along right MTG (red line) emerged in tandem (i.e., spatially covarying) with the associative model fit (green line), while the perceptual model fit (blue line) diverged from both. This was statistically evident in a peak location analysis: The location of individual participants' relational category model peak ($M = 19.81$, $SE = 2.05$, $CI [15.6966, 23.9145]$) was on average the same as the location of their associative model peak ($P = 1$), but was anteriorly shifted from the perceptual model peak ($t(35) = -3.13$, $P = 0.004$, $d = -0.65$). The slope of the relational category model fit across ROIs was positive but not significantly different from 0 ($M = 0.0009$, $SE = 0.0006$, $CI [-0.0003, 0.0020]$,

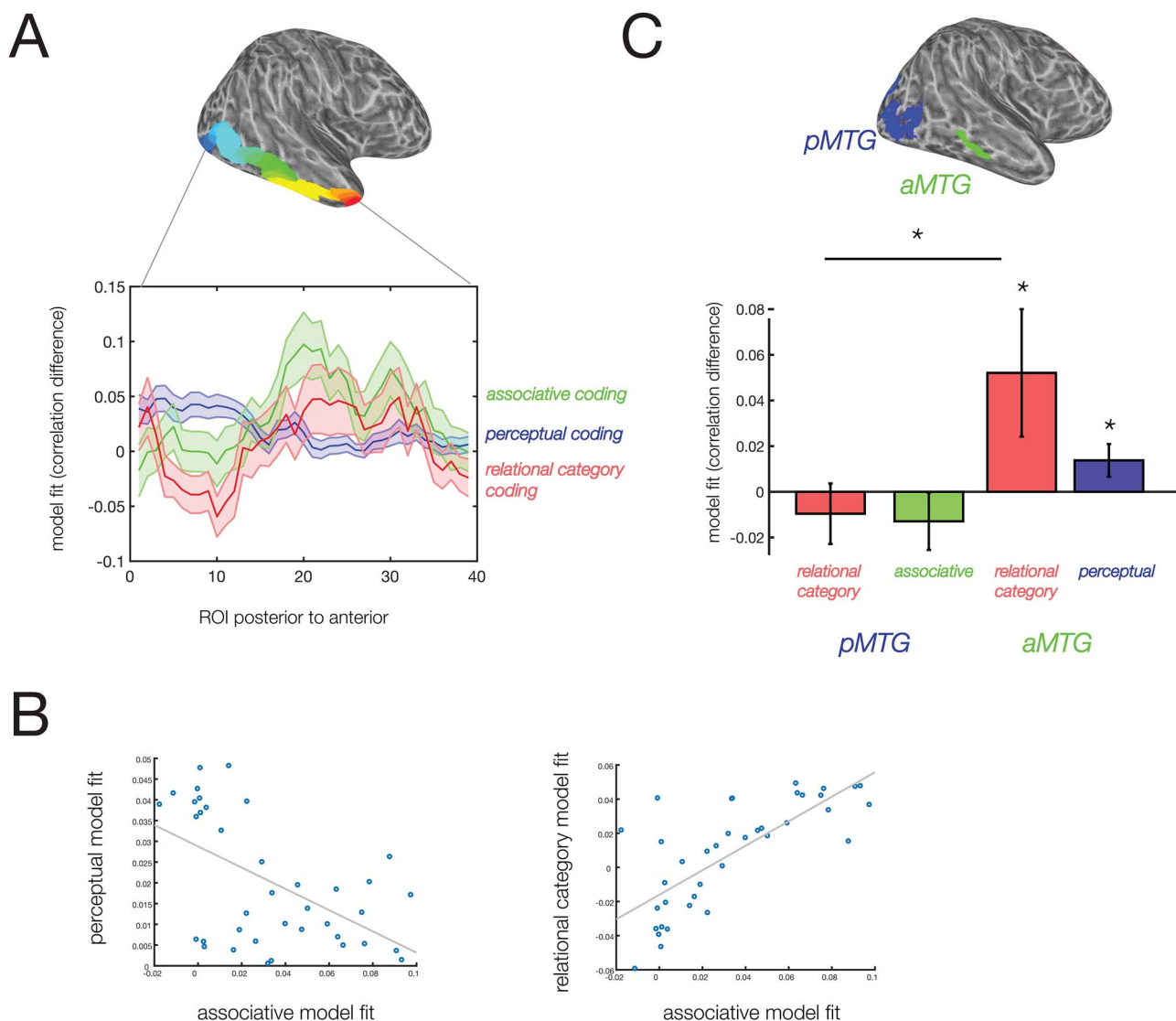


Figure 5. (A) Results of the vector of ROIs analysis along the posterior–anterior axis of right MTG. ROIs, color coded by number, are shown above; below, the fit of each of the three models (associative, perceptual, and relational category) is plotted against the index of each ROI. (B) Correlations between the associative and perceptual models (left) and associative and relational category models (right) across the mean fit values in each of the 39 ROIs in MTG. (C) Functional ROIs based on the associative model searchlight (aMTG) and perceptual model searchlight (pMTG), shown above, and the fit of the relational category model, associative model, and perceptual model, excluding the model used to define each ROI, below. * $P < 0.05$, uncorrected for multiple ROIs. This figure is available for viewing in full resolution at <https://osf.io/ydz8q/>.

$t(35) = 1.54$, $P = 0.134$), but it was significantly more positive than that of the perceptual model ($t(35) = -3.50$, $P = 0.001$). We further assessed this relationship by computing the correlation between the three models in terms of their fit (averaged across participants) along the 39 ROIs; these correlations are shown in **Figure 5B**. The correlation between perceptual model fit and associative model fit was strongly negative, $r(37) = -0.54$, $P < 0.001$, while the correlation between relational category model fit and associative model fit was strongly positive, $r(37) = 0.74$, $P < 0.001$, with a significant difference between them, $M = -1.29$, $z(37) = -6.36$, $P < 0.001$.

We performed a complementary test within individual participants, by taking each individual's vector of model fits across the 39 ROIs, computing the correlation among models, and testing these individual, fisher-corrected correlation values against

zero at the group level. Here we found a negative but nonsignificant correlation between associative and perceptual model fits, $M = -0.009$, $SE = 0.0580$, $t(35) < 1$, but a significant positive correlation between associative and relational category model fits, $M = 0.095$, $SE = 0.0459$, $CI [0.0027, 0.1864]$, $t(35) = 2.09$, $P = 0.044$, $d = 0.35$, though the difference between these two correlations was not significant, $t(35) = 1.50$, $P = 0.144$.

Although statistical tests of the associative and perceptual model fits relative to zero are biased, as their locations informed the ROI definition, the fit of the relational category model is independent. We found that the relational category model fit did not significantly exceed zero when correcting for the 39 ROIs, but that several anterior ROIs showed effects at uncorrected significance levels (ROI 31: $M = 0.049$, $SE = 0.027$, $CI [0.0043, Inf]$, $t(35) = 1.85$, $P = 0.036$, $d = 0.31$; ROI 33: $M = 0.040$,

SE=0.0213, CI [0.0049, Inf], $t(35)=1.92$, $P=0.031$, $d=0.32$, one-tailed tests). Despite not necessarily reliably exceeding zero, the close relationship between the magnitude of the relational category model fit and the associative model fit remains striking and informative and suggests an underlying systematicity to this signature.

A key feature of our design is that participants' task queried both forward and backward predictive relations (what comes next "and" before), so that participants would retrieve memory for both causes and effects in all conditions. However, one might argue that, despite these instructions, participants primarily retrieved forward predictions. If so, the correlation between relational category and associative coding could be driven by the fact that, in similar relational category blocks, participants predominantly retrieved the same event from memory (the effect, when seeing the cause). If these assumptions are true, then we should not see the same relative correlation effects between our three models when using trials showing the effect, random, and rare effects. We tested this idea. Despite the reduction in the amount of data included, we found a consistent set of results: Relational category coding was negatively correlated across ROIs with the perceptual model, $r(37)=-0.32$, $P=0.048$, but still positively correlated with the associative model, $r(37)=0.69$, $P<0.001$. Peak locations along the right MTG significantly diverged between relational category and perceptual coding effects, $M=5.69$, $t(35)=2.26$, $P=0.03$, but did not diverge between relational category and associative coding effects, $M=-1.67$, $P=0.500$. However, we did not see effects in individual subject correlations. Overall, however, this suggests that the correspondence between associative and relational category effects across right MTG is not driven only by the retrieval of a particular event during cause trials, but by the entire predictive pattern across the events in each condition.

In summary, we found evidence that relational category and specific associative coding are related in terms of their location along right MTG, while perceptual coding diverges from both, particularly at the group level (i.e., in terms of which models are reliable across subjects). In individual participants, relational category coding was anterior in peak location to perceptual coding, but overlapping with associative coding, and was correlated in fit across ROIs only with associative coding (though not significantly more so than with perceptual coding). When considering the cross-subject reliability of model fits across ROIs, we found that ROIs with stronger relational coding were more likely to have stronger associative coding, but less likely to have strong perceptual coding. Altogether, these data suggest that specific associative coding emerges spatially in tandem with increased relational category coding, but diverges from perceptual coding, along right MTG.

Functional ROIs

To additionally characterize the relationship among the three similarity models, we defined functional ROIs using the whole-brain searchlight results from the associative and perceptual models, by taking the significant group clusters and identifying individual participant ROIs within those clusters (see Functional ROI Definition). We always tested the two models not used to define the ROIs to assure independence. Analyses to measure model fit in these ROIs followed the same procedure as before.

Specifically, we used the whole-brain-corrected associative coding searchlight MVPA results to define ROIs in right anterior right MTG (aMTG), right prefrontal cortex (PFC), left PFC, left

intraparietal sulcus (IPS), left medial prefrontal cortex (MPFC), and right precuneus (PC); the clusters are shown in Figure 4A, and individual ROIs were defined within these clusters (see Methods). Perceptual coding searchlight results were similarly used to define right posterior right MTG (pMTG); this group-level cluster is shown in Figure 4B.

As shown in Figure 5C, in aMTG, we found significant fits for the relational category model, $M=0.052$, SE=0.027, CI [0.0056, Inf], $t(35)=1.89$, $P=0.033$ one-tailed, $d=0.32$, and the perceptual model, $M=0.013$, SE=0.007, CI [0.0019, Inf], $t(35)=1.96$, $P=0.029$ one-tailed, $d=0.33$, with no difference between them, $t(35)=1.34$, $P=0.189$. Thus, aMTG contained some perceptual information about the events, along with relational information. It should be noted however that these fits, relative to 0, do not survive correction for the testing of multiple functional ROIs and so should be interpreted cautiously.

Although the associative coding model was used to define the ROI and thus cannot be tested here, the vector of ROIs results described above (and shown in Figure 5A) established that associative coding is strongest in anterior parts of MTG relative to posterior parts, and the whole-brain contrast between these models (Fig. 6) shows that an anterior MTG cluster shows significantly stronger associative than perceptual coding. It should be noted that none of the ROIs in the vector were identical to the aMTG functional ROI, so minor differences between these are not contradictory.

More importantly, however, relational information was not predictive of patterns in pMTG, despite substantial perceptual coding. In pMTG, neither the associative nor the relational category model had significant fits (relational category model, $M=-0.0096$, SE=0.0133, CI [-0.0317, Inf], with $t(35)=-0.73$, $P=0.766$; associative model, $M=-0.0129$, SE=0.0125, CI [-0.0338, Inf], $t(35)=-1.05$, $P=0.845$). In-line with the vector of ROIs analysis, the fit of the relational category model was significantly stronger in aMTG than in pMTG, $M=-0.0096$, SE=0.0133, CI [-0.0317, Inf], $t(35)=2.62$, $P=0.013$, $d=0.47$, the latter of which was not significant, $t<1$. Overall, these analyses suggest that aMTG contains information about relational properties, both specific and (probably) general, as well as about the perceptual properties of events, while pMTG showed evidence only of perceptual information and substantially less associative information of any kind.

A different way to consider the relationship among models is to look at the correlation among model fits across participants within these functional ROIs. Here, we did not find that participants with higher associative model fits in aMTG had higher relational category model fits in this area, $r(34)=-0.05$, $P=0.774$. Instead, we saw a marginal correlation between associative model and perceptual model fits, $r(34)=0.30$, $P=0.074$, though this correlation was not significantly higher ($z(34)=1.46$, $P=0.144$). We did not see significant correlations among models in pMTG. The contrast between these results and those of the vector of ROIs analysis suggests that the relationship between models manifests specifically at the level of large-scale organization: that associative and relational category coding are both more likely to be found anteriorly, rather than posteriorly, while perceptual coding is more likely to be found in posteriorly, rather than anteriorly. Within these ROIs, however, model fit relationships across participants do not follow the same pattern and could simply reflect that participants who paid more attention to the events had higher fits for perceptual and associative models, while relational category models had some independent variance.

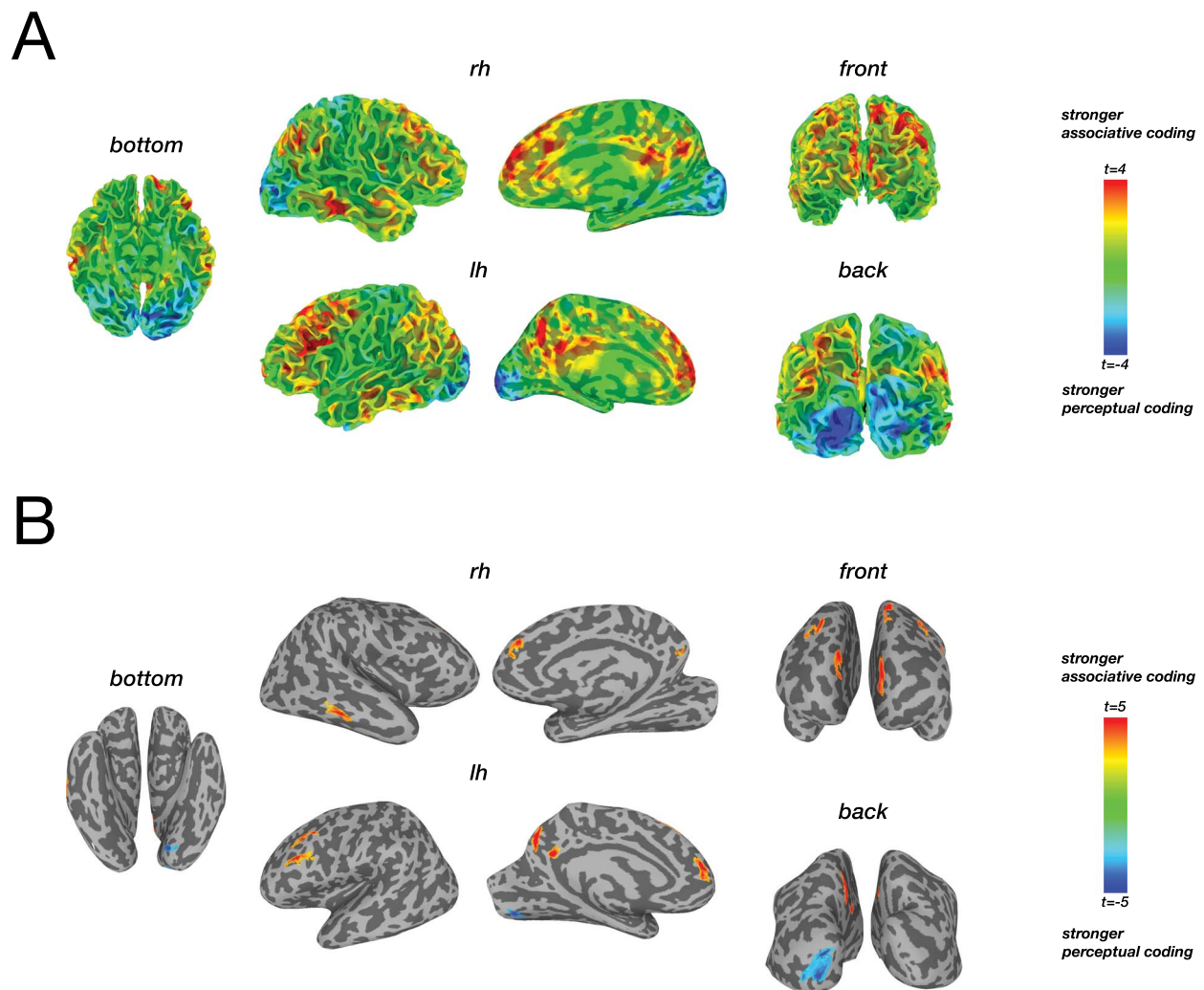


Figure 6. Whole-brain plots illustrating t-values for the comparison of associative model fits relative to perceptual model fits across participants. Hot colors show areas that had relatively stronger associative model fits, while cool colors show areas showing relatively stronger perceptual model fits. Figure A is for illustrative purposes; Figure 6B shows significance-thresholded results. This figure is available for viewing in full resolution at <https://osf.io/2wdyp/>.

In the remaining associative coding ROIs, we found significant fits of the perceptual model in left PC, left MPFC, left LPFC, and left IPS (Supplementary Table S3). Thus, even though these areas did not have strong enough perceptual model effects to appear in searchlight findings, most associative areas did contain information about the perceptual properties of the events. However, we did not find significant fits of the relational category model in any ROI (all $t < 1$). Thus, the pattern of divergence and overlap that we found in right MTG may be specific to this area.

Finally, we did not find any significant correlations between our behavioral measures (of accuracy or categorization) and model fits in any ROI, perhaps because participants were close to ceiling on these measures by design and thus exhibited limited range.

Discussion

We investigated the neural mechanisms supporting long-term predictive memory by probing three kinds of representations that the brain might simultaneously exhibit when processing

visual events: their visual features (perceptual coding), memory of their specific predictive relations (associative coding), and their generalized relational categories (relational category coding). We aimed to better understand the relationship between these kinds of representations in order to resolve two puzzles: 1) how the brain simultaneously encodes the distinctions between visually different events while also representing the predictive relations among them, and 2) how it might build generalizable representations of predictive relations across distinct contexts.

We found a diverse set of areas representing specific predictive relations (associative coding), but these areas were not the ones showing the strongest evidence of perceptual coding. Perceptual coding was found to be strongest in posterior aspects of the temporal lobe, including posterior MTG, an area known to be important for processing dynamic stimuli (Beauchamp et al. 2002). However, anteriorly along the right MTG, the neural response to an event began to increasingly resemble its visually distinct associate and decreasingly resemble its visual matches. The influence of predictive knowledge was negatively correlated with the strength of perceptual coding along the posterior–anterior axis of right MTG, and the peak locations

of these effects were reliably divergent in individual subjects (Fig. 5).

In whole-brain contrasts, we found that perceptual coding and associative coding differentially characterized a number of regions across the cortex: Associative coding was relatively stronger in anterior MTG and other areas anatomically resembling the default mode network (Figs 4 and 6). Perceptual coding was relatively stronger only in the most posterior aspects of the temporal lobe, areas likely involved in lower-level visual processes. Midanterior temporal areas showed significant perceptual coding (Fig. 4B) but did not exceed the strength of associative coding (Fig. 6), consistent with a gradual transition in the strength of these measures (Fig. 5).

In contrast to the divergence with perceptual coding, relational category coding increased spatially in tandem with associative coding along the right MTG and peaked at the same location, leading to greater relational category information in anterior than posterior right MTG. This suggests that posterior and anterior MTG encode different information about the same events: pMTG reflects only their perceptual properties, while aMTG additionally reflects memory of their predictive relations, including the way these relations generalize across variation (here, the participating objects). pMTG lacked this information despite the fact that retrieving predictive knowledge was highly task-relevant. In summary, along right MTG, as responses become more reflective of specific predictive relations, they also become more reflective of relational categories and less reflective of perceptual features.

These findings help resolve the puzzles we raised earlier by showing that visual discrimination and predictive relation retrieval are handled by distinct parts of the cortex, enabling the brain to accomplish both functions distinctly. This implies that prior work showing associative coding in the temporal lobe—most of it, from neurophysiology work in macaques (Miyashita 1988; Sakai and Miyashita 1991; Higuchi and Miyashita 1996; Erickson and Desimone 1999; Messinger et al. 2001; Naya et al. 2003)—might similarly have reflected areas that are not those which most strongly represent visual features (or vice versa). This is not necessarily clear from their anatomical locations, given the differences between humans and monkeys and across experiments.

Conversely, our findings also imply that generalized representations of relational categories and representations of specific relations both rely on common parts of right MTG, perhaps helping the brain build the former from the latter. Other regions exhibiting associative coding (apart from right MTG) did not exhibit information about relational categories. This suggests that right MTG may have a specialized role in bringing representations of events further away from their visual attributes and closer toward generalized, relational representations. This capacity is suggestive of a key role of this region in event understanding.

MTG as a Core Locus of Event Memory

Our findings are in-line with other work implicating various parts of MTG in event memory. Among other areas, MTG exhibits correlated patterns between watching a movie clip and then recalling it later and correlated activity between participants recalling the same movie (Chen et al. 2017). MTG tends to appear as part of a network of regions including the medial and lateral prefrontal cortex, angular gyrus, and precuneus, some of which appeared in our associative coding results also. These

regions form part of the default mode network, thought to be critical to both prospection and memory (Buckner et al. 2008), and related posterior-medial network, thought to be critical to encoding episodic memories in terms of relations among items and events (Ranganath and Ritchey 2012; Ranganath and Hsieh 2016). Recently, similar areas in lateral prefrontal cortex, precuneus, and MTG were implicated in inferring the abstract structure of a set of predictive relations (Tomov et al. 2018). However, MTG has been relatively underexplored relative to other parts of this network, and it is notable that we found it to have a unique representational signature among them.

Work on action and event knowledge offers some insight into why MTG might have had a particularly important role in our task. MTG is the most consistent area to show selective responses to retrieving action knowledge from memory (Martin et al. 1995; Kable et al. 2002, 2005; Phillips et al. 2002; Perini et al. 2014; Leshinskaya, Wurm and Caramazza (in press)). Nearby parts of MTG show selective responses to verbs (Bedny et al. 2008, 2011; Peelen et al. 2012) and/or event nouns (Bedny et al. 2013). These responses tend to be posterior and/or superior to our aMTG ROI, but it is likely that they fall somewhere within the perception-to-memory gradient we observe.

Finally, we did not see effects of relational categories in the region of left STG reported by Frankland and Greene (2015), where they identified patterns distinguishing sentences in which two objects were related in the same way (truck hitting the ball) or in the opposite way (ball hitting the truck). This could be seen as analogous to our manipulation of whether an object caused an event or reacted to the same event. However, we probed memory representations (the typical role of an object as represented in semantic memory) rather than the extraction of these events from syntactic information in the context of the task. Our findings suggest that it is possible that that memory-based relational categories diverge from those embedded in syntax, with left-lateralized STG specialized for relations as conveyed by syntax in particular. A direct comparison of these functions would be a fascinating topic for future research.

Neural Organization of Long-Term Associative Memory

The neural locus of long-term associative memory and its principles of organization are hardly settled. Prior work using similar paradigms, in which associative memory is probed following a separate, prior learning task, has identified associative coding signatures in the medial temporal lobes (Schapiro et al. 2012; Hindy et al. 2016; Garvert et al. 2017). In contrast, we failed to find strong evidence of associative coding in medial temporal lobes, finding stronger effects elsewhere. These prior experiments differed from ours in one important way: They used a substantially shorter delay between learning and scanning, about 1 day. It is well established that the importance of the hippocampus in associative memory declines with time, due to the effects of consolidation (Tse et al. 2007; Winocur et al. 2007; Yamashita et al. 2009). Indeed, consolidation research can show dramatic differences in memory signatures following delays of 24 h compared with delays of 30 days (Bontempi et al. 1999; Richards et al. 2014). We thus believe that our findings do not challenge prior results showing the importance of the hippocampus in associative memory with relatively short delays, but add important evidence regarding the loci of relatively more consolidated long-term memory.

Even with regard to paradigms with longer delays, the neural organization of long-term associative memory remains

unsettled. For example, such work has found evidence of associative coding in various parts of the IT lobes but has either not examined other areas or did not report them (Miyashita 1988; Sakai and Miyashita 1991; Erickson and Desimone 1999; Senoussi et al. 2016). However, even with such results, it has been unclear whether the neural areas responsible for associative coding are those that simultaneously support visual discrimination—functions which seem to be at odds. We show a divergence between perceptual and associative functions at the large scale and argue that long-term memory of predictive relations is represented predominantly in areas associated with other aspects of prospection, memory, and event knowledge (though not to the exclusion of others, particularly at a fine scale). The pattern we observe also broadly coincides with an increased sensitivity to spatial relations in more anterior versus posterior temporal areas (Kaiser and Peelen 2017; Baldassano et al. 2017a).

The gradient we observed from posterior to anterior MTG also coincides with observations of similarly oriented gradients of “temporal integration windows,” that is, where correlations between subjects viewing the same movie are affected by scrambling event order at shorter versus longer timescales (Lerner et al. 2011; Baldassano et al. 2017b). If anterior areas track information across a longer period of time, they could be more influenced by predictive/associative history. However, in movie stimuli, longer timescales also convey more relational content (such as interactions among actors), another reason why our findings may coincide.

Conclusion

Overall, our findings help address several previous unknowns regarding long-term predictive memory. We argue that right anterior MTG is particularly specialized for representing which specific events are predictive of each other, as opposed to which are visually similar, and that it captures generalized relational similarity across distinct contexts. By pulling apart visual and relational similarity in this way, and enabling generalization, MTG plays a pivotal role in event memory and understanding. More broadly, our findings illustrate a functional divergence between cortical areas representing events in terms of their remembered predictive relations versus their visual properties.

Funding

National Institutes of Health (P30 NS45839 to the Center for Functional Neuroimaging at the University of Pennsylvania, and R01EY021717, R01DC015359, and R01DC009209 to S.L.T.-S.).

Notes

We would like to thank Cristina Leon, Mira Bajaj, and Jennifer Stiso for assistance with programming, data collection, and analysis. We also thank Brice Kuhl, Christopher Honey, and Janice Chen for helpful discussion of these findings.

Conflict of Interest

None declared.

References

- Baldassano C, Beck DM, Fei-Fei L. 2017a. Human-object interactions are more than the sum of their parts. *Cereb Cortex*. 27(3):2276–2288.
- Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. 2017b. Discovering event structure in continuous narrative perception and memory. *Neuron*. 95(3):709–721.
- Beauchamp MS, Lee KE, Haxby JV, Martin A. 2002. Parallel visual motion processing streams for manipulable objects and human movements. *Neuron*. 34:149–159.
- Bedny M, Caramazza A, Grossman E, Pascual-Leone A, Saxe R. 2008. Concepts are more than percepts: the case of action verbs. *J Neurosci*. 28(44):11347–11353.
- Bedny M, Caramazza A, Pascual-Leone A, Saxe R. 2011. Typical neural representations of action verbs develop without vision. *Cereb Cortex*. 22(2):286–293.
- Bedny M, Dravida S, Saxe R. 2013. Shindigs, brunches, and rodeos: the neural basis of event words. *Cogn Affect Behav Neurosci*. 14(3):891–901.
- Buckner RL, Andrews-Hanna JR, Schacter DL. 2008. The brain's default network: anatomy, function, and relevance to disease. *Ann N Y Acad Sci*. 1124(March):1–38.
- Chen J, Leong YC, Honey CJ, Yong CH, Norman KA, Hasson U. 2017. Shared memories reveal shared structure in neural activity across individuals. *Nat Neurosci*. 20(1):115–125.
- Christie S, Gentner D. 2010. Where hypotheses come from: learning new relations by structural alignment. *J Cogn Dev*. 11(3):356–373.
- Corral D, Jones M. 2014. The effects of relational structure on analogical learning. *Cognition*. 132(3):280–300.
- Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res*. 29(3):162–173.
- Erickson CA, Desimone R. 1999. Responses of macaque perirhinal neurons during and after visual stimulus association learning. *J Neurosci*. 19(23):10404–10416.
- Favila SE, Chanales AJH, Kuhl BA. 2016. Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat Commun*. 7(11066):1–10.
- Fischl B, Sereno MI, Tootell RI, Dale AM. 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp*. 8(4):272–284.
- Frankland SM, Greene JD. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proc Natl Acad Sci USA*. 112(37):11732–11737.
- Garvert MM, Dolan RJ, Behrens TEJ. 2017. A map of abstract relational knowledge in the human hippocampal – entorhinal cortex. *ELife*. 6:1–20.
- Gentner D. 1983. Structure mapping: a theoretical framework for analogy. *Cognitive Sci*. 7(2).
- Goldstone RL, Medin DL, Gentner D. 1991. Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychol*. 23(2):222–262.
- Goldwater MB, Gentner D. 2015. On the acquisition of abstract knowledge: structural alignment and explication in learning causal system categories. *Cognition*. 137:137–153.
- Gopnik A, Meltzoff AN. 1997. *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
- Haxby JV, Gobbini IM, Furey ML. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*. 293:2425–2430.

- Higuchi S, Miyashita Y. 1996. Formation of mnemonic neuronal responses to visual paired associates in inferotemporal cortex is impaired by perirhinal and entorhinal lesions. *Proc Natl Acad Sci USA*. 93(2):739–743.
- Hindy NC, Ng FY, Turk-Browne NB. 2016. Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat Neurosci*. 19(5):665–667.
- Jones M, Love BC. 2007. Beyond common features: the role of roles in determining similarity. *Cognitive Psychol*. 55(3):196–231.
- Kable JW, Kan IP, Wilson A, Thompson-Schill SL, Chatterjee A. 2005. Conceptual representations of action in the lateral temporal cortex. *J Cogn Neurosci*. 17(12):1855–1870.
- Kable JW, Lease-Spellmeyer J, Chatterjee A. 2002. Neural substrates of action event knowledge. *J Cogn Neurosci*. 14(5):795–805.
- Kaiser D, Peelen MV. 2017. Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *Neuroimage*. 169:334–341.
- Kemp C, Tenenbaum JB, Niyogi S, Griffiths TL. 2010. A probabilistic model of theory formation. *Cognition*. 114(2):165–196.
- Kok P, Turk-Browne NB. 2018. Associative prediction of visual shape in the hippocampus. *J Neurosci*. 38(31):6888–6899.
- Konkle T, Caramazza A. 2013. Tripartite organization of the ventral stream by animacy and object size. *J Neurosci*. 33(25):10235–10242.
- Kuhl BA, Chun MM. 2014. Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *J Neurosci*. 34(23):8051–8060.
- Lerner Y, Honey CJ, Silbert LJ, Hasson U. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J Neurosci*. 31(8):2906–2915.
- Leshinskaya A, Thompson-Schill SL. (2019). From the structure of experience to concepts of structure: how the concept “cause” applies to streams of events. *Journal of Experimental Psychology-General*. 148(4):619–643.
- Leshinskaya A, Wurm MF, Caramazza A. (in press). Concepts of Actions and their Objects. In M Gazzaniga, GR Mangun, D Poeppel. *The Cognitive Neurosciences*, 6th edition, 757–765.
- Long NM, Lee H, Kuhl BA. 2016. Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes. *J Neurosci*. 36(50):1850–1816.
- Markman AB, Gentner D. 1993. Structural alignment during similarity comparisons. *Cognitive Psychol*. 25(4):431–467.
- Markman AB, Stilwell CH. 2001. Role-governed categories. *J Exp Theor Artif Intell*. 13(4):329–358.
- Martin A, Haxby JV, Lalonde FM, Wiggs CL, Lalonde FM, Ungerleider LG. 1995. Discrete cortical regions associated with knowledge of color and knowledge of action. *Science*. 270(5233):102–105.
- Messinger A, Squire LR, Zola SM, Albright TD. 2001. Neuronal representations of stimulus associations develop in the temporal lobe during learning. *Proc Natl Acad Sci USA*. 98(21):12239–12244.
- Miyashita Y. 1988. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*. 335(27):817–820.
- Mumford S. 1998. *Dispositions*. Oxford: Oxford University Press.
- Naya Y, Yoshida M, Miyashita Y. 2003. Forward processing of long-term associative memory in monkey inferotemporal cortex. *J Neurosci*. 23(7):2861–2871.
- Oosterhof NN, Wiestler T, Diedrichsen J. 2014. Surfing: a matlab toolbox for surface-based voxel selection. Retrieved from <http://surfing.sourceforge.net>.
- Oosterhof NN, Wiestler T, Downing PE, Diedrichsen J. 2011. A comparison of volume-based and surface-based multi-voxel pattern analysis. *NeuroImage*. 56(2):593–600.
- Oosterhof NN, Wiggett AJ, Diedrichsen J, Tipper SP, Downing PE. 2010b. Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. *J Neurophysiol*. 104(2):1077–1089.
- Peelen MV, Romagno D, Caramazza A. 2012. Independent representations of verbs and actions in left lateral temporal cortex. *J Cogn Neurosci*. 24(10):2096–2107.
- Perini F, Caramazza A, Peelen MV. 2014. Left occipitotemporal cortex contributes to the discrimination of tool-associated hand actions: fMRI and TMS evidence. *Front Hum Neurosci*. 8:1–10.
- Phillips JA, Noppeney U, Humphreys GW, Price CJ. 2002. Can segregation within the semantic system account for category-specific deficits? *Brain*. 125:2067–2080.
- Pinker S. 1989. *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: MIT Press.
- Polyn SM, Natu VS, Cohen JD, Norman KA. 2005. Category-specific cortical activity precedes retrieval during memory search. *Science*. 310(5756):1963–1966.
- Ranganath C, Hsieh L-T. 2016. The hippocampus: a special place for time. *Ann N Y Acad Sci*. 1369(1):93–110.
- Ranganath C, Ritchey M. 2012. Two cortical systems for memory-guided behaviour. *Nat Rev Neurosci*. 13(10):713–726.
- Reddy L, Poncet M, Self MW, Peters JC, Douw L, Van Dellen E, Claus S, Reijneveld JS, Baayen JC, Roelfsema PR. 2015. Learning of anticipatory responses in single neurons of the human medial temporal lobe. *Nat Commun*. 6:1–8.
- Richards BA, Xia F, Santoro A, Husse J, Woodin MA, Josselyn SA, Frankland PW. 2014. Patterns across multiple memories are identified over time. *Nature Neuroscience*. 17(7):981–986.
- Sakai K, Miyashita Y. 1991. Neural organization for the long-term memory of paired associates. *Nature*. 354(6349):152–155.
- Schapiro AC, Kustner LV, Turk-Browne NB. 2012. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biol*. 22(17):1622–1627.
- Senoussi M, Berry I, VanRullen R, Reddy L. 2016. Multivoxel object representations in adult human visual cortex are flexible: an associative learning study. *J Cogn Neurosci*. 28(6):852–868.
- Stuhlmüller A, Tenenbaum JB, Goodman ND. 2010. Learning structured generative concepts. *Proc Annu Conf Cogn Sci Soc*. 32:2296–2301.
- Tomov MS, Dorfman HM, Gershman SJ. 2018. Neural computations underlying causal structure learning. *J Neurosci*. 38(32):7143–7157.
- Tse D, Langston RF, Kakeyama M, Bethus I, Spooner PA, Wood ER, Witter MP, Morris RGM. 2007. Schemas and memory consolidation. *Science*. 316(5821):76–82.
- Turk-Browne NB, Scholl BJ, Johnson MK, Chun MM. 2010. Implicit perceptual anticipation triggered by statistical learning. *J Neurosci*. 30(33):11177–11187.
- Wang J, Cherkassky VL, Yang Y, Chang KK, Vargas R, Diana N, Just MA. 2016. Identifying thematic roles from fmri-measured neural representations. *Cogn Neuropsychol*. 33(3–4):257–264.

- Winocur G, Moscovitch M, Sekeres M. 2007. Memory consolidation or transformation: context manipulation and hippocampal representations of memory. *Nat Neurosci.* 10(5): 555–557.
- Yamashita K-I, Hirose S, Kunitatsu A, Aoki S, Chikazoe J, Masutani Y et al. 2009. Formation of long-term memory representation in human temporal cortex related to pictorial paired associates. *J Neurosci.* 29(33):10335–10340.
- Zeithamova D, Dominick AL, Preston AR. 2012. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron.* 75(1): 168–179.