ANNUAL REVIEWS

Annual Review of Psychology

Concepts and Compositionality: In Search of the Brain's Language of Thought

Steven M. Frankland[1] and Joshua D. Greene[2]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08544, USA; email: steven.frankland@princeton.edu

[2]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA; email: jgreene@wjh.harvard.edu

Annu. Rev. Psychol. 2020. 71:17.1–17.31

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-122216-011829

Keywords

compositionality, conceptual combination, language of thought, default mode network, grid cells, artificial intelligence

Abstract

Imagine Genghis Khan, Aretha Franklin, and the Cleveland Cavaliers performing an opera on Maui. This silly sentence makes a serious point: As humans, we can flexibly generate and comprehend an unbounded number of complex ideas. Little is known, however, about how our brains accomplish this. Here we assemble clues from disparate areas of cognitive neuroscience, integrating recent research on language, memory, episodic simulation, and computational models of high-level cognition. Our review is framed by Fodor's classic language of thought hypothesis, according to which our minds employ an amodal, language-like system for combining and recombining simple concepts to form more complex thoughts. Here, we highlight emerging work on combinatorial processes in the brain and consider this work's relation to the language of thought. We review evidence for distinct, but complementary, contributions of map-like representations in subregions of the default mode network and sentence-like representations of conceptual relations in regions of the temporal and prefrontal cortex.

Contents

BACKGROUND AND OVERVIEW: COMPOSITIONALITY AND THE LANGUAGE OF THOUGHT

In *A Treatise of Human Nature*, eighteenth-century Scottish philosopher David Hume imagines the city of New Jerusalem, "whose pavement is gold and walls are rubies" [Hume 2003 (1738), p. 8; see also Fodor 2003). Like Hume, we can imagine cities with *gold pavement* and *rubied walls* because our minds can combine simpler ideas (*pavement*, *gold*, *streets*, *made of*) to think about unfamiliar objects, places, and events. This ability to systematically construct new thoughts out of old ideas is a hallmark of human cognition. More than 250 years after Hume's writing, the biological and computational bases of that ability remain elusive: How does the brain reuse bits of prior knowledge to generate new thoughts?

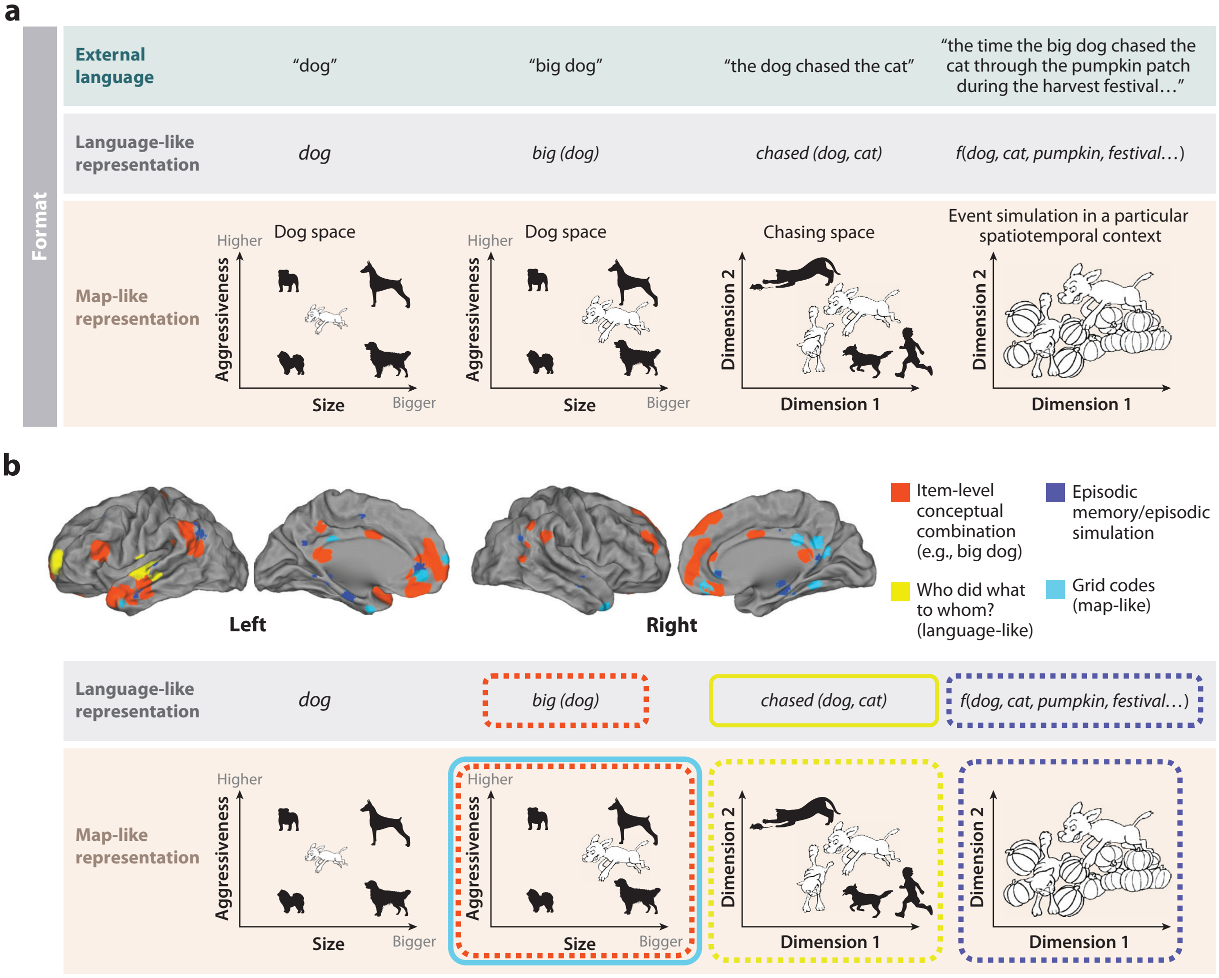
Natural languages and other symbolic systems (such as mathematical and computer programming languages) are powerful because they are compositional, making "infinite use of finite means" through the flexible recombination of simpler parts (Von Humboldt, cited in Chomsky 2006, p. 15). The principle of compositionality holds that the meaning of a complex expression is a function of the meaning of its parts and the way in which they are combined (Frege 1980). In a natural language, these complex expressions (e.g., phrases or sentences) are built by taking simpler symbolic representations from a reusable set (e.g., words or morphemes) and applying abstract syntactic operations to differentially assemble them. Some cognitive scientists have proposed that the same basic principles underlie the generativity of nonlinguistic thought in humans (Fodor 1975, Fodor & Pylyshyn 1988; see also Jackendoff 1992) and perhaps in nonhuman species (Fodor & Pylyshyn 1988; but see Penn et al. 2008). According to this view, the mind composes complex thoughts by reassembling conceptual representations expressed in an internal symbolic language, that is, a language of thought (LoT) (Fodor 1975, Fodor & Pylyshyn 1988). (See Newell 1980 for a classic explication of symbolic architectures; see Piantadosi 2011, Goodman et al. 2014, and Piantadosi & Jacobs 2016 for recent reapplications of the LoT hypothesis to psychology.)

A cognitive architecture that flexibly recombines representational parts (e.g., concepts) using abstract procedures has a number of appealing theoretical properties (Fodor & Pylyshyn 1988). First, it explains the productivity of thought, or why the human mind can generate and understand an astronomically large number of ideas, such as the idea of *Abraham Lincoln being serenaded* *by nineteen purple penguins on the deck of an aircraft carrier while sailing down the Nile*. It also explains the systematicity of thought, or why, for example, anyone who can think about *a doctor needing an accountant* can also think about an *accountant needing a doctor*. According to LoT, these involve the same component concepts (*needs*, *doctor*, *accountant*) and combinatorial procedures for mapping components to roles [*needs* (x, y)], but with the mapping from concepts to roles reversed (cf. Johnson 2004 for skepticism regarding the notion of systematicity). This account contrasts with a phrasebook model of language and thought. A tourist relying on a phrasebook might know how to say *Where is the train station?* and *Please take me to the hospital* but not know how to say *Please take me to the train station*. The LoT hypothesis thus explains why we, as thinkers, are not like hapless tourists bound by a limited phrasebook of ideas.

For these reasons, LoT has played an important role in the philosophy of mind and in some branches of cognitive science. However, there has been a considerable divide between the central ideas of LoT and empirical and theoretical work in cognitive and computational neuroscience. While this divide partly reflects an indifference to biological implementation among the chief proponents of LoT (Fodor 1975, Fodor & Pylyshyn 1988), it also has more substantive sources. First, it owes to the conceptual gap between the operation of the "dry" silicon circuits that serve as the canonical physical vehicle for implementing symbolic architectures and the "wet" circuits of the human brain, characterized by probabilistic neural spiking and experience-dependent synaptic modification. How could the brain possibly implement or approximate symbolic architectures? This conceptual gap has narrowed over the last three decades, as theorists have proposed sophisticated mechanisms for approximating the flexible construction of complex, structured representations in biologically inspired systems (Smolensky 1990, Plate 1995, Hummel et al. 2004, Rougier et al. 2005, Van der Velde & de Kamps 2006, Doumas et al. 2008, Hayworth 2012, Eliasmith 2013, Kriete et al. 2013). Second, the gap owes partly to the relative dearth of empirical work on the neural representation of complex, structured content—the chief explanatory virtue of the LoT. There is a substantial literature on neural representations of objects and actions (e.g., Martin & Chao 2001, Thompson-Schill 2003, Martin 2007, Patterson et al. 2007, Mitchell et al. 2008, Bedny et al. 2008, Mahon & Caramazza 2009, Binder et al. 2009, Ralph et al. 2017, Wurm & Caramazza 2019), but less is known about the combinatorial processing of these representations. Here we review and synthesize an emerging body of work on the brain's high-level combinatorial processes. By "combinatorial" we mean processes that can take as inputs a vast range of preexisting independent representations and produce more complex representations. For example, one might combine *ruby* and *walls* to imagine *rubied walls*, or combine *doctor*, *accountant*, and *needs* to encode either that *the doctor needs an accountant* or *the accountant needs a doctor*. Combinatorial processes, in addition to generating representations with phrase- and sentence-level meaning (e.g., describing events or states of affairs), can also generate concepts themselves, (e.g., an *aunt* is a *sister* of a *parent*), which may then serve as units within more complex compositions.

Here, we use the LoT as a guiding framework and consider how such a representational system may map onto the macro-level representational strategies in the human brain (**Figure 1**). We review recent neuroimaging work on conceptual combination (e.g., Hume's combining *gold* and *pavement* to imagine *gold pavement*) and find that, across studies, combinatorial processes engage a common set of brain regions, typically housed throughout the brain's default mode network (DMN). We relate this work on conceptual combination to more familiar processes attributed to the DMN, such as episodic memory, semantic memory, prospection, social cognition, and event comprehension. We then focus on particular encoding schemes for complex representations, including recent work on the grid-like encodings in the DMN, which may reflect an abstract relational structure that is shared across domains. We then consider how these grid-like representations differ from more arbitrary and explicit structured representations, such as encoding who did what to whom in an event and differentiating, for example, the meaning of *the gold pavement is brighter than the rubied walls* from that of *the rubied walls are brighter than the gold pavement*. Finally, we review recent advances in the computational modeling of high-level cognition that we believe will be important for a mechanistic understanding of how the brain builds complex, structured thoughts.

Drawing on these findings, we sketch a set of hypotheses concerning the brain's high-level strategies for composing complex concepts and relate them to the LoT hypothesis. We suggest that the DMN retrieves and integrates stored structural knowledge to construct thoughts in a context-dependent reference frame. We suggest that the DMN's operations are distinct from those devoted to particular perceptual modalities and shared across semantic domains and input modalities (perception, language comprehension, and internally driven memory retrieval). However, the representational format may be more like a map (Tolman 1948, Camp 2007, Behrens et al. 2018) than the sentence-like representations favored by the classic LoT hypothesis (Fodor 1975). (See also Anderson 1978, Pearson & Kosslyn 2015 for theoretical work on various



(*Caption appears on following page*)

Figure 1 (*Figure appears on preceding page*)

A language of thought (LoT) framework for the neural representation of complex ideas. (*a*) Following Fodor (1975), we propose that the brain composes complex ideas by employing reusable amodal representations, but we suggest that these come in complementary formats. As in the classic LoT hypothesis, we propose that some representations roughly follow the structure of external, natural language (*first row*), with symbol-like structures representing the values of abstract semantic variables (*second row*). Such representations can serve as instructions for generating richer conceptual representations using the knowledge implicitly encoded in cognitive maps (*third row*). (*b*) Recent empirical studies of conceptual combination (*red*) reveal effects throughout the default mode network (DMN) (Graves et al. 2010; Baron & Osherson 2011; Bemis & Pylkkänen 2011, 2013a; Pallier et al. 2011; Barron et al. 2013; Coutanche & Thompson-Schill 2014; Price et al. 2015) (coordinates from Bemis & Pylkkänen 2011, 2013a estimated from visualized MEG data). Doeller et al. (2010) and Constantinescu et al. (2016) provide evidence for the use of grid-like codes in the navigation of physical and conceptual spaces (respectively), with effects observed throughout the DMN (*light blue*). Results for conceptual combination and grid-like representation align roughly with the DMN regions implicated in episodic memory/simulation (Benoit & Schacter 2015) (*dark blue*). Frankland & Greene (2015, 2019) provide evidence for the use of reusable symbol-like representations in encoding the relational structure of events (i.e., Who did what to whom?) in the superior temporal gyrus and rostral prefrontal cortex (*yellow*). Solid lines reflect evidence for specific representational formats, while dotted lines reflect more speculative interpretations. Neural markers are 6-mm spheres centered on focal coordinates, warped to the cortical surface. Figure art credits: bull dog (**http://clipart-library.com/bulldog-silhouette-cliparts.html**), doberman (**http://getdrawings.com/doberman-pinscher-silhouette#doberman-pinscher-silhouette-12.jpg**), lapdog (**https://www.kissclipart.com/chow-chow-silhouette-clipart-chow-chow-finnish-lap-3hhn44/**), golden retriever (**http://www.supercoloring.com/silhouettes/golden-retriever**), dog-cat-pumpkins (**http://dogcoloringpages.org/dog-chases-cat-coloring-page.html**), cat-mouse (**http://clipart-library.com/clipart/752221.htm**), boy-dog (**http://www.clipartguide.com/_pages/0511-1008-0522-4315.html**).

representational formats in the context of mental imagery.) A map encodes structured, relational information between locations by assuming some isomorphism (paradigmatically spatial) between the representing and represented domain (Camp 2007; J.D. Cohen, unpublished manuscript). Intriguingly, grid-like representations in the DMN encode structural information present across semantic domains (Constantinescu et al. 2016; Aronov et al. 2017; Garvert et al. 2017; J.D. Cohen, unpublished manuscript). This abstract map-like representation in the DMN likely differs, however, from the representations in the fronto-parietal control network, which define operations over abstract variables corresponding to explicit semantic relations. The prefrontal cortex (PFC) has long been known to flexibly represent task-relevant content and to coordinate the execution of sequential programs based on current goals (Fuster 1997, Miller & Cohen 2001, Rougier et al. 2005, Sigala et al. 2008, Kriete et al. 2013). Here, we highlight two particular regions of the frontal and temporal lobes that may serve as representational way stations between highly flexible PFC regions and context-dependent coordinate spaces (map-like representations) in the DMN. These intermediate stops may be more akin to the sentence-like representations envisaged in the original LoT hypothesis, encoding explicit categorical relations at different levels of abstraction (Frankland & Greene 2015, 2019). We take the former system, grounded in the dorsolateral prefrontal cortex (DLPFC), to furnish the instructions for possible thoughts, and the latter, grounded in the DMN, to serve as the canvas on which those thoughts are drawn.

CONCEPTUAL COMBINATION

Although conceptual combination has been a topic of long-standing interest in cognitive psychology (Smith & Osherson 1984, Medin & Shoben 1988, Murphy 1988, Smith et al. 1988, Gagné & Shoben 1997, Wisniewski 1997, Estes & Glucksberg 2000, Wisniewski & Middleton 2002), it has only recently received attention in cognitive neuroscience. We first review recent neuroimaging work that has identified regions supporting conceptual combination (e.g., Hume's combination of *rubied* and *walls* to imagine *rubied walls*, and *gold* and *pavement* to imagine *golden pavement*). In the sections that follow, we consider how such regions might encode the relevant information.

The brain regions most reliably implicated in conceptual combination appear to lie within the DMN[1] (Baron et al. 2010; Graves et al. 2010; Baron & Osherson 2011; Bemis & Pylkkänen 2011, 2013a; Forgács et al. 2012; Molinaro et al. 2012, 2015; Pylkkänen et al. 2014; Price et al. 2015, 2016; Westerlund et al. 2015). Specifically, regions of the medial prefrontal cortex (mPFC) (Graves et al. 2010, Bemis & Pylkkänen 2011, Forgács et al. 2012, Bemis & Pylkkänen 2013b, Pylkkänen et al. 2014), posterior cingulate cortex (PCC) (Graves et al. 2010, Baron & Osherson 2011), angular gyrus (Graves et al. 2010; Bemis & Pylkkänen 2013b; Boylan et al. 2015; Price et al. 2015, 2016), and, prominently, the lateral mid-anterior temporal cortex (Baron et al. 2010, Baron & Osherson 2011, Bemis & Pylkkänen 2011, Coutanche & Thompson-Schill 2014, Westerlund et al. 2015) have been implicated in conceptual combination across measurement techniques, semantic domains, and input modalities. There appears to be no single brain region responsible for conceptual combination, but the brain regions that constitute the DMN appear to have a distinctive role (**Figure 1*b***, *orange*).

In an early functional magnetic resonance imaging (fMRI) study of conceptual combination (Graves et al. 2010), participants read pairs of nouns (noun-noun combinations) that formed semantically plausible combinations when presented in one order (e.g., *lake house*) but less plausible combinations when reversed (e.g., *house lake*). Graves and colleagues contrasted the mean BOLD signal evoked by these two conditions (plausible versus implausible) and found that plausible noun-noun orderings (*lake house*) produce greater activation than implausible combinations (*house lake*) in three DMN regions: the PCC, the dorsomedial prefrontal cortex (dmPFC), and a right-lateralized part of the angular gyrus. Forgács et al. (2012) find the same broad set of DMN regions (PCC, mPFC, and bilateral angular gyrus) in a study of German noun-noun combinations.

Noun-noun combinations are distinctive in that their representation is not determined by syntax alone, and therefore they require inferences about how the words' referents are most likely related (e.g., Is a *lake house* more likely to be a typical house *located by* a lake or a highly atypical house *made out of* a lake?) (Murphy 1988, Gagné & Shoben 1997, Wisniewski 1997). However, cases of conceptual combination that are more closely guided by syntax have also been found to elicit effects in the same DMN regions: mPFC (Bemis & Pylkkänen 2011, 2013b; Pylkkänen et al. 2014), angular gyrus (Bemis & Pylkkänen 2013a; Price et al. 2015, 2016), and PCC (Baron & Osherson 2011). For example, Price et al. (2015, 2016) provide converging evidence for bilateral angular gyrus involvement in adjective-noun conceptual combination, whereby meaningful pairs (e.g., *plaid jacket*) induced greater BOLD signal in the left angular gyrus than the presentation of less meaningful pairs (*moss pony*). Consistent with this, in patients with degenerative disease, gray matter density in both right and left angular gyri predicts behavioral impairment on conceptual combination but not single-word conceptual tasks (Price et al. 2015). Moreover, transcranial direct current stimulation (tDCS) of the left (but not right) angular gyrus facilitates faster meaningful/nonmeaningful judgments in the same task (Price et al. 2016). Pallier et al. (2011) and Boylan et al. (2015) also identify combinatorial effects in regions near the angular gyrus and the temporoparietal junction (TPJ) more broadly.

[1] The cortical regions that have to date been implicated in conceptual combination may ultimately be discovered to be anatomically and functionally distinct from those that exhibit higher levels of activity during rest, the feature that initially defined the DMN (Shulman et al. 1997, Raichle et al. 2001) (see, for example, Humphreys & Ralph 2014, Humphreys et al. 2015). Nonetheless, here we find it useful to refer to this set of regions throughout as DMN or DMN-like regions, given their broad anatomical correspondence to this well-studied network. Moreover, for present purposes, we will tend to treat the DMN as a single network, but recent work has characterized different subnetworks of the DMN (Andrews-Hanna et al. 2010, Christoff et al. 2016, Braga & Buckner 2017).

Studies using magnetoencephalography (MEG) provide convergent evidence that both the mPFC (Bemis & Pylkkänen 2011, 2013b; Pylkkänen et al. 2014) and the angular gyrus (Bemis & Pylkkänen 2013a) contribute to adjective-noun combination. For example, Bemis & Pylkkänen (2011) examined responses to adjective-noun combinations (*red boat*), comparing responses to noun-noun lists (*cup*, *boat*) and non-pronounceable string-noun lists (*xqz boat*). Of these, only the former (*red boat*) involves constructing a representation in which one word-meaning modifies the representation of the other. Thus, brain regions supporting composition should exhibit greater activity to *red boat* than to *cup*, *boat*, and *xqz boat*. Across studies, the authors have reported this signature in the mPFC (Bemis & Pylkkänen 2011, 2013b; Pylkkänen et al. 2014), angular gyrus (Bemis & Pylkkänen 2013a), and, most reliably, in the left mid-anterior temporal lobe (Bemis & Pylkkänen 2011, Westerlund & Pylkkänen 2014, Ziegler & Pylkkänen 2016). This set of temporal regions may correspond to the DMN-sub3 described by Christoff et al. (2016).

MEG studies examining the fine-grained time course of adjective-noun composition indicate that the left temporal region and the DMN regions (angular gyrus and mPFC) play complementary roles in adjective-noun composition. During reading comprehension tasks, Bemis & Pylkkänen (2011, 2013a) have found that the temporal lobe exhibits combinatorial effects early in the comprehension process (<250 ms following stimulus presentation). By contrast, the combinatorial effects in both mPFC (Bemis & Pylkkänen 2011) and angular gyrus (Bemis & Pylkkänen 2013a) are more delayed (∼400 ms and ∼350 ms, respectively, for reading, and 100–150 ms slower for auditory comprehension). The order of effects depends on whether the task is language production or comprehension. Roughly speaking, production begins with an understanding of what is to be communicated and ends with (the motor production of) a specific natural language expression. By contrast, comprehension begins with (the sensory encoding of) a specific natural language expression and ends with an understanding of what was communicated. For production, Pylkkänen et al. (2014) report that combinatorial effects in mPFC either precede (experiment 1) or co-occur with (experiment 2) combinatorial effects in the anterior temporal lobe.

What, then, are these different regions contributing to conceptual combination? Pylkkänen and colleagues have found that the left mid-anterior temporal lobe supports a restricted set of compositional operations. We highlight two key aspects. First, lateral anterior activity is driven by the magnitude of the content update (Westerlund & Pylkkänen 2014, Zhang & Pylkkänen 2015, Ziegler & Pylkkänen 2016). For example, although *canoe* (a subordinate category) elicits a stronger response in the left anterior temporal lobe (lATL) than *boat* (a basic-level category), the combination *blue boat* elicits a stronger response in the lATL (and TPJ regions) than *blue canoe* (Westerlund & Pylkkänen 2014). This effect is broadly consistent with Smith et al.'s (1988) classic model of conceptual combination. Their model starts with an abstract prototype representation of a concept (e.g., *boat*), modifies a particular dimension of that concept (e.g., the color), and raises the diagnosticity (importance) of the relevant dimension (e.g., the importance of color in differentiating a *blue boat* from a *red boat*). Imagining a *blue canoe* requires only modifying the color of a particular, pre-specified type of *boat* (a *canoe*), while imagining a *blue boat* requires first selecting a particular type of *boat* from the vast space of possibilities (e.g., *canoe*, *yacht*, *pirate ship*), and then modifying the color, thus requiring a greater traversal of conceptual space.

Second, the lATL effect depends on binding the properties to a particular individual, rather than describing an ensemble of objects (Del Prato & Pylkkänen 2014, Poortman & Pylkkänen 2016). For example, compare *the trees were green and big* to *the trees were small and big* (Poortman & Pylkkänen 2016). In the former case, one might imagine that each tree shares both features, requiring feature binding within an individual object representation. In the latter, one likely imagines an ensemble of trees, some of which are small and some of which are large. Poortman and Pylkkänen found that lATL responds more to within-object feature binding (intersective conjunctions, e.g., *the trees were green and big*) than to across-object conjunction (collective conjunctions, e.g., *the trees were small and big*). Unlike the lATL, the right anterior temporal lobe (rATL) selectively responded to collective but not intersective conjunctions.

This anatomical division of labor is consistent with the within-object combinatorial feature effects reported by Coutanche & Thompson-Schill (2014). Here, participants were instructed to look for objects drawn from four food categories (*lime*, *orange*, *celery*, and *carrot*), crossing shape and color features. As expected, the authors found that posterior temporal regions coded for the independent dimensions of shape (*th* versus *round*) and color (*orange* versus *green*). However, critically, they identified a medial region of the mid-anterior superior temporal sulcus (STS) as representing the conjunction of these features (e.g., a pattern unique to *green & round*). This conjunctive representation is consistent with the binding of separate object features within a token object representation. Taken together with the results from Pylkkänen and colleagues (Bemis & Pylkkänen 2011, 2013a; Pylkkänen et al. 2014; Westerlund et al. 2015; see also Baron & Osherson 2011), these findings suggest a broader role for the mid-anterior temporal cortex in integrating and modifying representations in the LoT, used both to construct word-level representations and to reuse word-level representations to construct compound- and phrase-level meaning.

Pallier et al. (2011) provide further evidence for the accumulation of information within a token representation in the lATL as well as in the mid-STS and TPJ. Here, participants read 12-word sequences. Critically, the syntactic units within these sequences varied in size from 1 to 12 words. For example, the sequence *looking ahead important task who dies his dog few holes they write* contained only two-word syntactic units (*looking ahead*, *important task*, etc.). At the other extreme, participants read, *I believe that you should accept the proposal of your new associate*. In this case, each of the 12 words contributed to the same semantic representation. The authors searched for brain regions in which activity monotonically tracked constituent size, and they identified the left temporal pole, mid-STS, and the left TPJ. However, unlike classic language areas in the lateral inferior frontal cortex and posterior superior temporal gyrus (STG), this effect was only observed when the constituents had conceptual content. No such monotonic effect was observed in the temporal pole, anterior STS, and TPJ for jabberwocky-infused word sequences such as *I tosieve that you should begept the tropufal of your tew viroate* (see also Hagoort & Indefrey 2014).

The representations described above are complex because, despite their differences, all result from the integration of multiple conceptual pieces. Compositionality, however, requires something more: The same representations must be reused across different instances within an architecture that supports their flexible recombination. Do these reported signatures reflect such reuse? It is possible that these examples, (e.g., *lake house*) are encoded without any internal semantic structure, as in a one-word name or a random bar code, and are simply retrieved from memory, phrasebook-wise, as a single unit. Compositionality, however, requires that the same component representations in *lake house* should be used for different, but related, combinations as in *lake station* and *mountain house*.

Barron et al. (2013) use fMRI adaptation to address this issue. They ask whether the representation of a novel, complex food is built using the simpler representations of the component foods (e.g., using *tea* and *jelly* to construct *tea jelly*, i.e., tea-flavored jelly). fMRI adaptation exploits the tendency for neural populations to diminish in firing rate when reactivated in response to a second stimulus, thus providing an index of reuse. The authors found that mPFC showed an attenuated BOLD response to combinations such as *tea jelly* if participants had just imagined either of the components independently (i.e., *tea* or *jelly* alone), consistent with a representational reuse of components in the composition.

Multi-voxel pattern analysis has revealed related evidence of conceptual reuse in the mid-anterior STS and PCC (Baron & Osherson 2011). Baron and colleagues (Baron et al. 2010, Baron & Osherson 2011) report that the pattern of BOLD signals in the mid-anterior STS and PCC evoked by complex concepts (e.g., *old woman*) can be modeled as a multiplication of the representation of the constituent concepts, consistent with Smolensky's (1990) tensor product model of binding. For example, to predict the pattern of activity for *old woman*, the authors separately estimated the independent contributions of *old* and *woman* across uses and then multiplied those patterns to generate the predicted activity to the conjunct *old woman*. Baron & Osherson (2011) find a region of the mid-anterior STS and a region of the PCC that are predicted by multiplicative combination. They find that mPFC patterns evoked by conceptual combinations can also be modeled as a systematic function of the patterns evoked by the components, but using an additive (*old* + *woman*) rather than a multiplicative (*old* × *woman*) model. Such an additive signal in the mPFC could reflect the retrieval of multiple memories, which are maintained in an independent form (e.g., by separate neural populations situated within the same voxel range, or by two successive retrieval steps, one for *old* and one for *woman*), with the interactions between these co-activated concepts computed elsewhere in the brain.

In theoretical work, Hummel and colleagues have emphasized the importance of maintaining orthogonal components to preserve the independence of roles (in this context, adjective and noun) and fillers (*old* and *woman*), and they have pursued computational models that might predict such an additive signal (Hummel & Holyoak 1997, Hummel et al. 2004, Doumas & Hummel 2012). In these particular models, the dynamic binding of roles and fillers is thought to be implemented by temporal synchrony (or asynchrony) between the component representations: A role and a filler are encoded separately but bound (or unbound) through the modulation of dynamically varying temporal phase relations (Shastri & Ajjanagadde 1993, Hummel & Holyoak 1997, Doumas et al. 2008, Doumas & Hummel 2012). (See Rabagliati et al. 2017 for the discovery of behavioral signatures in a conceptual combination task that were predicted by binding-by-asynchrony models; but see Shadlen & Movshon 1999 and O'Reilly & Busby 2002 for skepticism regarding the biological plausibility of binding through synchrony.) Moreover, synchrony is not the only mechanism for binding variables and values in a way that enables independence. The independent components of a tensor product (multiplications) can be recovered as long as the two vectors are orthogonal (Smolensky 1990, Plate 1995), and pointer-based architectures achieve role-filler independence as well (Eliasmith 2013, Kriete et al. 2013). We revisit recent computational work on variable binding in the final section on computational modeling.

Regardless of the physical mechanism of binding, we note that independent (i.e., role-filler independence) (Hummel et al. 2004) and integrated representations are both necessary for human-level understanding and generalization. The principle of compositionality assumes that there is some invariance in the contribution of *doctor* across the sentences (*a*) *the doctor needed an accountant* and (*b*) *the accountant needed a doctor.* However, a full understanding of a particular combination must flexibly estimate the interactions between the component parts: *Gray hair* is more like *white hair* than *black hair*, while *gray clouds* are more like *black clouds* than *white clouds* (Medin & Shoben 1988; see also Musslick et al. 2017 for computational work on the trade-offs between separate and shared representation in other domains). We revisit below the issue of context sensitivity (see the section titled Amodal Content) as well as the issue of representational reuse (see the section titled Who Did What to Whom).

In sum, these findings provide evidence that parts of the mPFC, angular gyrus, left mid-anterior superior temporal cortex, and PCC contribute to the construction of complex conceptual representations out of simpler parts. Further, some of these regions (mid-STS, PCC, and mPFC) show signs of reusing representations across conceptual combinations in a way that is prima facie consistent with compositionality. This adds conceptual combination to a list of cognitive processes attributed to the DMN (Raichle et al. 2001, Buckner & Carroll 2007, Hassabis & Maguire 2007, Hassabis et al. 2007, Buckner et al. 2008, Schacter et al. 2008, Christoff et al. 2016). How, then, do the above findings fit in with our emerging understanding of the DMN and its representational capacities? And in what sense might the DMN implement a language of thought?

REPRESENTATION IN THE DEFAULT MODE NETWORK

The DMN was originally defined as a set of regions preferentially engaged during rest (Shulman et al. 1997, Mazoyer et al. 2001, Raichle et al. 2001). Later work implicated the DMN in introspective processes, such as self-referential thought (Johnson et al. 2002, D'Argembeau et al. 2005) and mind wandering (Christoff et al. 2009, 2016). A series of influential reviews—Buckner & Carroll (2007), Buckner et al. (2008), Schacter et al. (2008)—broadened the conception of the DMN by implicating it in functions such as episodic memory (Buckner et al. 2005, Vincent et al. 2006), prospection (Addis et al. 2007, Schacter et al. 2007, Andrews-Hanna et al. 2010, Spreng et al. 2010, Schacter et al. 2012), navigation (Maguire et al. 1998, Spreng et al. 2010), and theory of mind (Saxe et al. 2004, Amodio & Frith 2006). Integrating these findings, Buckner et al. (2008, p. 2) argue that the DMN's core function is the flexible construction of "self-relevant mental explorations–simulations" of possible events (see also Hassabis & Maguire (2007) on the idea of scene construction in the DMN) (**Figure 1*b***, *dark blue*). Simulating an event, the authors observe, requires disengaging attention from the current environment and retrieving information from long-term memory. Consistent with this profile, the DMN exhibits preferential anatomical and functional connectivity with the brain's declarative memory systems (paradigmatically, the hippocampus and medial temporal lobe) rather than with sensory/motor cortices (Buckner et al. 2008).

Conceptually, we assume that such simulations require the retrieval and use of structured conceptual knowledge: One must incorporate probabilistic knowledge about what is likely to happen, given the features of the simulated content, and embed those in a particular spatiotemporal context (see also Irish et al. 2012, Irish & Piguet 2013 for empirical work supporting the relationship between semantic memory and episodic simulation, and Binder et al. 2009 on shared underlying brain regions). Although the LoT hypothesis comes in stronger and weaker forms, the central idea is that computations supporting high-level cognition operate over tokens expressed in an amodal internal language. These tokens are constructed by applying combinatorial procedures to reusable representations drawn from memory such that the syntax (structure) of the representation is distinct from its semantics (content).[2] For example, the thoughts *the accountant needs a doctor* and *the doctor needs an accountant* both involve the abstract structure [*needs* (x, y)] that can be accessed separately of any particular instance of the *needs* relation, such as *the dog needs a bath*. Although conceptual knowledge about the relationships between entities and events (i.e., semantic memory) is likely stored in the synaptic connections throughout the cerebral cortex, we assume that the tokens in a LoT over which computations operate are time-varying neural activation states. Thus, within a LoT framework, the episodic simulations in the DMN would involve the construction and updating of token representations (activation states) over which further computations may occur.

[2] In the strongest construal of the LoT hypothesis (Fodor 1975, Fodor & Pylyshyn 1988), the higher-order computations that operate on these representations are only sensitive to the structure (syntax) of the representations and not to their content (semantics). Though this syntax sensitivity is a central part of the classical, symbolic view of the mind, it can limit the types of cognitive operations that the mind can execute (logical inferences like P & Q → P being paradigmatic cases). We discuss the ways in which current approaches avoid this syntactic constraint to model many aspects of cognition (Eliasmith 2013; Kriete et al. 2013; Graves et al. 2014, 2016; Lake et al. 2015).

Our suggested mapping of aspects of the LoT onto the DMN thus raises a number of important questions. In the next two sections, first, we consider evidence for an amodal representational format in the DMN not specific to a particular input modality or content domain, and, second, we consider representational formats that may separate structure and content.

Amodal Content

In an influential meta-analysis of fMRI and positron emission topography (PET) data, Binder et al. (2009) found that DMN regions track semantic processing in language across a host of tasks. Here, we highlight recent work that has added to our understanding of the representational function of the DMN. Specifically, we highlight work that has found (*a*) that patterns of activity in DMN regions carry information about current, behaviorally relevant conceptual content (Fairhall & Caramazza 2013, Baldassano et al. 2017, Chen et al. 2017, Zadbood et al. 2017), and (*b*) that DMN involvement in conceptual processing is likely not restricted to language, as it also supports representations drawn from other input modalities, such as naturalistic videos (Baldassano et al. 2017, Chen et al. 2017, Zadbood et al. 2017) and images (Fairhall & Caramazza 2013), consistent with the existence of amodal representations.

Consider, first, the naturalistic movie-viewing paradigm of Hasson and colleagues (Baldassano et al. 2017, Chen et al. 2017, Zadbood et al. 2017), whereby participants watch continuous videos of an unfamiliar TV show while undergoing fMRI. Chen et al. (2017) found that particular scenes in the show evoked reliable, discriminable patterns of activity in the DMN and that these patterns were shared across participants. Moreover, the same scene-specific patterns observed during event encoding were later re-instantiated during the unconstrained free recall of the events in the video. Zadbood et al. (2017) replicated Chen et al.'s (2017) encoding/recall finding, but additionally, they scanned a separate group of participants while listening to verbal descriptions of the events from the videos generated by the observers of the original video. They found that hearing verbal descriptions of these events evoked similar patterns of neural activity in the PCC, mPFC, and angular gyrus compared to observing these events (see also Baldassano et al. 2017). These studies thus provide evidence that the DMN may not be restricted to the consideration of nonpresent events (Buckner & Carroll 2007, Schacter et al. 2008) or fictions (Hassabis & Maguire 2007), but that it may also support the encoding of current events (given that the events shown on video may be fictional, but their display is a real event). Moreover, these cross-modal results are consistent with an important aspect of the LoT: Conceptual content is shared across modalities (watching movies versus listening to descriptions of those movies).

It remains unclear, however, what aspects of the observed events are enabling this classification. The movie is comprised of complex events that vary in the particular individuals, objects, scene statistics, and emotional content that are present. Any of these might promote successful scene identification. However, data provided by Fairhall & Caramazza (2013) suggest that the representational capacity of the DMN extends beyond scenes to particular object categories (e.g., *mammals*, *birds*, *tools*, *fruits*, or *clothes*) (see also Devereux et al. 2013, Crittenden et al. 2015, Smith et al. 2018). They found that patterns in DMN regions (PCC, angular gyrus, and dmPFC) as well as non-DMN regions [DLPFC, posterior middle temporal gyrus (MTG), and inferior temporal cortex support classification of the categories of visually presented words and images, as well as cross-modal generalization of category information across words and images. That is, classifiers trained to identify which word is present (e.g., *apple*) can also classify pictures (e.g., a picture of an apple), suggesting a shared underlying conceptual representation across modalities.

Thus, it appears that DMN regions contribute to the representation of conceptual content and do so across different input modalities, such as words (Binder et al. 2009, Fairhall & Caramazza

2013, Zadbood et al. 2017), images (Fairhall & Caramazza 2013), and videos (Baldassano et al. 2017, Chen et al. 2017, Zadbood et al. 2017). Critically, these regions encode aspects of current stimuli (Hasson et al. 2008, Lerner et al. 2011, Fairhall & Caramazza 2013, Constantinescu et al. 2016, Baldassano et al. 2017, Chen et al. 2017) and current task context (Crittenden et al. 2015, Smith et al. 2018), and not just past, future, or hypothetical realities.

Thus, the DMN appears to support the utilization of rich, structured knowledge across modalities. However, the DMN likely does not store conceptual knowledge about particular domains. Instead, knowledge of specific semantic content appears to be represented in domain-specific (or, at least, highly specialized) cortical areas. There are, for example, anatomical dissociations based on sensory-motor features (Thompson-Schill 2003, Patterson et al. 2007), more abstract features such as object size (Konkle & Caramazza 2013), and even more abstract conceptual domains, such as animate versus inanimate entities (Mahon et al. 2009, Connolly et al. 2012). (See also Mitchell et al. 2008, Huth et al. 2016 on distributed semantic representations, and Leshinskaya & Caramazza 2015 for the representation of abstract function knowledge in the anterior parietal cortex). The overall organization of these semantic domains remains a topic of debate (Martin 2007, Mahon & Caramazza 2009). These domain-specific representations are then believed to depend on amodal conceptual hubs in the anterior temporal lobe for their integration (Patterson et al. 2007, Ralph et al. 2017). Notably, recent evidence has also implicated the posterior MTG in the amodal representation of conceptual information across objects (Devereux et al. 2013, Fairhall & Caramazza 2013) and actions (Wurm & Caramazza 2019). The particular representational content of these amodal conceptual representations appears to differ from the apparently amodal representations in the DMN, however.

Margulies et al. (2016) suggest that the DMN sits atop a representational hierarchy, representing the current landscape at the greatest level of abstraction. Consistent with this, Hasson and colleagues (Hasson et al. 2008, Lerner et al. 2011, Honey et al. 2012, Regev et al. 2013) found that DMN regions accumulated information more slowly over time than sensory regions and could thus integrate information about the recent temporal context. Moreover, Crittenden et al. (2015) and Smith et al. (2018) provided evidence that DMN response magnitude is modulated by task switches, suggesting that the DMN supports context representations (Smith et al. 2018): When the representational context changes, DMN neurons must update the dimensions of variation they are currently encoding, causing an observable increase in BOLD signal. The regions of the DMN, unlike those specialized for particular semantic domains, are thus perhaps best understood as a kind of amodal conceptual canvas dependent on the recent context rather than as a repository of semantic knowledge.

But what could context-dependent representations in the DMN have to do with conceptual combination? To a first approximation, one might think that combining concepts (e.g., Hume's *rubied walls*) is just computing the intersection of two or more sets (e.g., the set of things made of *rubies* and the set of things that are *walls*). However, it has long been known that a word's semantic contribution is not rigidly fixed across different uses, but instead it interacts with those of the other words in the combination. Combinatorial processes are therefore also inferences, requiring estimates of which possible states would best explain the use of an expression (see, for example, Medin & Shoben 1988, Murphy 1988, Hampton 1997, Wisniewski 1997, Estes & Glucksberg 2000, Wisniewski & Middleton 2002). This, of course, requires careful attention to context—both background context and the context that the focal semantic items create for each other. A *small car* is bigger than a *big grapefruit*, and, as noted earlier, *gray hair* is more like *white hair* than *black hair*, while *gray clouds* are more like *black clouds* than *white clouds* (Medin & Shoben 1988). As these examples suggest, understanding even simple sentences requires rich knowledge about the interaction between types of objects and events, and such understanding would be difficult to implement in a serial symbolic processor such as a classical Turing machine (Fodor 2001, Pinker 2005), which relies on syntactic form for inference.[3] Such understanding seems to require rapid "all things considered" probabilistic inference rather than the application of simple syntactic rules. It is therefore interesting to consider whether a shared set of global processes for inference to the best explanation may account for the common engagement of DMN regions in social cognitive tasks, such as theory of mind (Saxe et al. 2004, Amodio & Frith 2006, Tamir et al. 2015), as these require highly abstract prior knowledge to rapidly infer which mental states are most likely to have caused an observed action. The relationship between the DMN's representational contributions to theory of mind, episodic simulation, and conceptual combination is thus an important topic for future empirical work. This task is all the more pressing given the recent finding that the DMN may in fact be composed of distinct, spatially interdigitated networks that are typically blurred in group analyses (Braga & Buckner 2017).

Grid Cells and Conceptual Maps

How, then, might abstract, structured knowledge about the current context be represented? Recent research has highlighted grid cells as a kind of neural encoding mechanism that appears to exist throughout DMN regions and to support an amodal representation of the current context. Specifically, grid cells are believed to provide a low-dimensional, abstract representation of the metric structure shared across spatial environments (Hafting et al. 2005, Dordek et al. 2016, Stachenfeld et al. 2017), enabling spatial navigation. They have been most extensively studied in the rodent medial entorhinal cortex (MEC) but appear to be common across mammals, including humans (Jacobs et al. 2013, Doeller et al. 2010, Bellmund et al. 2016, Constantinescu et al. 2016, Horner et al. 2016). In humans, their functional signatures have been observed in a broader set of DMN regions, such as mPFC (Doeller et al. 2010, Constantinescu et al. 2016), PCC (Doeller et al. 2010, Constantinescu et al. 2016), and TPJ (Doeller et al. 2010, Constantinescu et al. 2016).

Rather than coding for a particular location, a grid cell fires at periodically spaced locations throughout an environment (Hafting et al. 2005), with responses arranged in equilateral triangular lattices, forming hexagonal structures. This periodic response differs from other neural coding schemes, such as classical population codes in which neurons are selectively tuned to particular values of a variable, as in orientation detectors in V1, and place cells in the hippocampus. Within the rodent MEC, grid cells are organized in discrete modules devoted to particular spatial frequencies (and aligned dorsal to ventral in the MEC), with different cells within a module representing different phases for each frequency. While place cells remap across environments, grid cells are relatively consistent, with their primary distortion being a reorientation in novel settings (Barry et al. 2007, Carpenter et al. 2015). Moreover, they are relatively insensitive to perceptual input (Hafting et al. 2005) and thus provide an abstract representation of the structure of environments that may differ in surface perceptual features.

A number of recent computational studies have found that grid-like representations can be derived by simply observing the similarity structure of one's experience in a spatial environment

[3] These findings are not incompatible with the existence of a central representational medium (a language) over which mental computations are performed, but they cast doubt on the most rigid versions of the LoT hypothesis, according to which high-level computations are solely restricted to rule-like operations over syntactic structure, as in classic Turing machines. Generative probabilistic models (see below) have proven to be useful high-level descriptors of how a particular representation is constructed, and they provide a valuable model of concepts more generally (though this is separate from the question of encoding that we discuss throughout). Readers are referred to Piantadosi (2011) and Goodman et al. (2014) for related ideas regarding a probabilistic LoT.

over time (given that nearby points are likely to be visited in close temporal succession), and by either factoring that stream of experience into a lower-dimensional representation (Dordek et al. 2016, Stachenfeld et al. 2017) or predicting which location one is likely to visit in the near future given one's current state (Cueva & Wei 2018, Whittington et al. 2018). Thus, they provide support for abstract map-like representations of a particular context or domain. Maps, unlike sentences, exhibit a spatial isomorphism between the content represented and the representation itself: Objects (and the locations they occupy) that are nearby in space are nearby on a 2D map, subject to some scale of variation (Camp 2007). However, map-like representations in the brain need not themselves be spatially laid out across the cortical sheet as in a topographic map of the domain, such as orientation detectors in V1. Instead, we consider these representations map-like if there exists an isomorphism between the similarity relations in the representational space (e.g., the hippocampal place cells and the factorization of their similarity relationships into grid codes) and the similarity relations of the represented content. Grid cells appear to exploit the similarity relations in hippocampal place cells to extract the underlying dimensions of variation in the space (e.g., the *x* and *y* axes of a 2D environment).

We highlight grid cells here because of recent work showing that grid-like representations are not limited to spatial navigation (Doeller et al. 2010), or even imagined navigation (Horner et al. 2016), but appear to be employed across representational domains, such as sound frequency (Aronov et al. 2017) and, more importantly for present purposes, abstract conceptual dimensions (Constantinescu et al. 2016). Although there has been direct evidence of grid cells in the human entorhinal cortex using intracranial recording (Jacobs et al. 2013), some of the most relevant work here comes from indirect evidence that is derived from an ingenious fMRI method developed by Doeller et al. (2010) and later adopted by Constantinescu et al. (2016) to study traversals through abstract conceptual space (Gärdenfors 2004). In this paradigm, Doeller et al. (2010) exploited the fact that grid cells consistently aligned to the external environment at a particular angle within individuals. Because of the hexagonal response structure of the grid cells, traversing the space in multiples of 60° from the grid angle produces greater grid-cell activity than moving at off-60° multiples. Remarkably, because this particular alignment is shared by all grid cells, these trajectories produce macro-level differences in aggregate BOLD signal that are detectable in humans. Doeller et al. (2010) observed this difference in the entorhinal cortex, mPFC, and PCC (DMN regions) during spatial navigation.

Constantinescu et al. (2016) adopted Doeller and colleagues' fMRI paradigm, but rather than navigating a 2D box, the subjects performed tasks that required navigating a 2D conceptual space. One axis of the space was the length of a bird's neck, and the other axis was the length of the bird's legs. Though images of the birds were not visually presented, the authors nonetheless observed the hexadirectional grid signature in the mPFC, PCC, and angular gyrus when participants were mentally transforming one neck-leg length combination into another. Thus, continuous traversal of abstract conceptual space appears to exploit grid-like representations, as in spatial navigation (see **Figure 1*b***, *light blue*).

This suggests an intriguing explanation for the involvement of DMN regions in tasks requiring conceptual combination. Forming a complex concept such as *red boat* requires modifying a preexisting representation: in this case, modifying a representation of a particular category of object (*boat*) along a particular dimension (color). The grid-like representations in the DMN suggest that these complex concepts may be encoded through the application of vector displacement procedures. In spatial navigation, these representations sum displacement signals over time to update the current state (position in space). However, rather than continuous summing of incremental movement in navigation, conceptual combination requires larger, immediate displacements in position. It is therefore unlikely that combining *gray*, *brown*, and *dog* to think of a *gray and brown dog* would produce a hexadirectional signal detectable with fMRI. Nevertheless, the DMN's involvement in conceptual combination may reflect the activation of representations in abstract conceptual spaces, defined with grid-like structure. When there is common structure across experiences (as in space), these representations may be reused to facilitate rapid computation and inference, and they may be updated based on the current context. More generally, the reuse of grid-like representations may enable computational operations originally evolved for spatial navigation to be applied to more abstract things (Gärdenfors 2004). Notably, spatial notions such as *distance* and *betweenness* apply naturally to concepts of all kinds (Bellmund et al. 2018), as when we say that *a fax machine is in between a photocopier and a phone* or that *a chair is closer to a loveseat than a lamp*.

These observations suggest a role for grid-like representations—and abstract conceptual maps more generally—within a system capable of compositional thought. As noted above, representing concepts within a spatial framework may support useful computations, such as computations of distance. In classic work, Shepard (1987) characterized the relationship between psychological distance and generalization across a number of representational dimensions (size, lightness, spectral hue, Morse code signals, etc.) and species (e.g., humans and pigeons). Shepard noted that in all domains, generalization decays as an exponential function of psychological distance to a familiar stimulus, suggesting the utility and ubiquity of the metric representation across domains. Perhaps map-like representations of space may support the development of new conceptual combinations through forms of interpolation and limited extrapolation. Representing photocopiers and telephones within a common conceptual space practically invites the invention of the fax machine, and it makes one wonder what forms of transportation might lie beyond airplanes and rockets.

In relevant computational modeling work, Webb and colleagues (T. Webb, S.M. Frankland, A.A. Petrov, R.C. O'Reilly & J.D. Cohen, unpublished manuscript) recently studied the interpolation and extrapolation capacity of neural networks on a relational reasoning task that required the determination of the direction and distance between two points (i.e., How far is *x* from *y*?). The authors showed that first transforming input representations into a canonical representational space preserves the relational structure in the input. Critically, this transformation enables not only successful interpolation to novel points within a familiar range but also extrapolation to novel regions of the space, held out of the network's training. This exceeds the generalization capacities of neural networks trained on other familiar types of input representation (e.g., place codes, rate codes), and it much more closely approximates the extrapolation ability of humans that is often lacking in artificial neural networks. Webb and colleagues define this procedure, which normalizes received inputs based on the structure of recently experienced stimuli, as "context normalization." This raises the possibility that the DMN may perform such context normalization and that context normalization may enable extrapolation from past experiences to structurally similar, but novel, cases.

What, then, are grid representations capable of encoding? In relevant theoretical work, Fiete et al. (2008) showed that grid cells exhibit a surprisingly large combinatorial capacity. Their encoding capacity grows exponentially with the number of cells, scaled by the number of distinct periods. They argued that grid-cell encodings are examples of "modulo" codes, like those used in residue number systems (RNS) applied in cryptography. In an RNS, each integer is expressed as the remainder of a division by a finite set of integers (called moduli). In the grid-cell case, this value is the remainder of phase divided by period for that module (i.e., distance between locations of maximum response). The computation of this remainder is the modulo operation. A given position may be represented by a population of cells that compute such remainders across phases and periods, providing a dense code of current position. RNS-based grid-like codes support simple algorithms for arithmetic ($a + b$, $a - b$, $a \times b$), which enables continuous position updating during movement.

However, despite its advantages, an RNS (and hence a grid code) has revealing limitations. Critically, an RNS representation discourages the use of simple algorithms for computing directional relations (a > b) (Garner 1959, Fiete et al. 2008). In order to compute these categorical relations in an RNS, it is conventional to translate from an RNS to a different representational system, such as a fixed-base system (Garner 1959) in which relative magnitude can be read off using a combination of positional information (syntax) and base value (content), as in decimal notation. For example, the statements that $5 > 2$, $25 < 52$, and $222 > 52$ are easily computable by general algorithms that integrate information about numeric position and content. Such explicit directional relations are difficult to compute in an RNS (here, a grid code) without simply memorizing a full table of inequalities. However, humans can generalize directional relations well beyond the scope of their experience. Explicit categorical relations may thus be difficult to compute in a system that consists only of grid cells.

Moreover, limits of spatial models appear in related domains of human behavior. For example, in classic work, Tversky (1977) notes that human similarity judgments fail to follow a number of properties that are axiomatic of metric spaces (and hence spatial representations, such as those implemented in grid cells). One such axiom is the symmetry of distances [$d(\mathrm{A}, \mathrm{B}) = d(\mathrm{B}, \mathrm{A})$]. To the extent that human similarity judgments depend on a spatial representation, these judgments should be symmetric as well, that is, A is as similar (far from) B as B is to A. However, they often are not. Consider Tversky's classic example: Humans tend to judge North Korea to be more similar to China than they judge China to be to North Korea. Thus, although the computation of distance in a metric space is order invariant, human similarity judgments are often order dependent. Tversky (1977) notes that linguistic statements of similarity are themselves directional: *A is like B* does not equal *B is like A*. The syntactic object (or referent) tends to be a salient, static point, whereas the subject (A) is defined with respect to that point. Compare, for example, *the president is like a child* to *a child is like the president*. Thus, human similarity judgments do not appear to be simple, context-independent distance computations but depend on the relative position of the objects to compare.

These explorations suggest a role for grid-like representations and cognitive maps within a LoT, but they also suggest that more is needed, both on empirical grounds (Tversky 1977) and theoretical grounds (Garner 1959, Fiete et al. 2008). Between the concept of a photocopier and the concept of a phone, there was a new concept, a new idea, just waiting to be born. But was there a location in one of our cognitive maps just waiting for the idea of *Abraham Lincoln being serenaded by nineteen purple penguins on the deck of an aircraft carrier while sailing down the Nile*? Perhaps. Alternatively, it may be that cognitive maps of the kind supported by the DMN and grid-like networks can only get someone—to put it in spatial terms—so far. Consistent with this, Gu et al. (2015) argue that the DMN facilitates the instantiation of neural patterns that are relatively easy to reach, while other networks, such as the frontoparietal network, push the brain into difficult-to-reach states. Some conceptual combinations, it seems, are more difficult to reach than others. It is unlikely, for example, that you would have wandered into our Abraham Lincoln example on your own. The DMN may be essential for retrieving and utilizing prior knowledge about structural dependencies between states to form a context, but it may need help reaching far-flung points in conceptual space. What gives us the ability to comprehend and generate arbitrary conceptual combinations involving explicit relations? To paint an unexpectedly new composition, the DMN may need special instructions.

WHO DID WHAT TO WHOM: A CASE STUDY OF EXPLICIT STRUCTURED RELATIONS

While map-like representations implicitly encode certain kinds of conceptual relations, such as context-specific similarity, they may be less useful for encoding explicit structure, especially involving asymmetric relations. Here, we address this question as we review recent empirical work on the representation of who did what to whom in response to descriptions of events. How, for example, does the brain construct distinct meanings involving the same elements, such as *the dog bit the man* and *the man bit the dog* (Pinker 1997)? In one sentence, the dog is the doer (the agent) and the man is the thing that has something done to it (the patient). In the other, the mapping of concepts to roles is reversed. Simply activating representations of the relevant concepts (*man*, *dog*, *bite*) is not enough to capture the meaning, nor is inferring what is most likely to be true in general (e.g., Do dogs bite men more often or the other way around?). Instead, these concepts must be bound to different roles, representing a particular relational position (e.g., Who did the biting in this case?).

Above we highlighted the systematic nature of human thought (Evans 1982, Fodor & Pylyshyn 1988): Anyone who can understand *the dog bit the man* can also understand *the man bit the dog*. Once again, the LoT explanation for this pattern is that these meanings are not encoded holistically in phrasebook fashion but instead use the same representations and combinatorial procedures. Fodor & Pylyshyn (1988, p. 13) take the mapping between an abstract LoT and the brain literally, writing that "the symbol structures…are assumed to correspond to real physical structures in the brain and the combinatorial structure of a representation is supposed to have a counterpart in the structural relations among physical properties of the brain." This is not the only possibility (e.g., Smolensky 1990, Doumas et al. 2008), but in our own research, we have taken seriously the idea that the LoT, if real, could have an observable structural basis in the brain.

In a series of studies (Frankland & Greene 2015, 2019), we targeted the neural representation of events in which an agent does something to a patient (**Figure 1*b***, *yellow*)**.** Participants read simple transitive sentences such as *the dog chased the cat* and *the cat chased the dog*, presented in either the active or passive voice, while undergoing fMRI. We identified a region of the left-mid superior temporal cortex (lmSTC) whose patterns of activity differentiate between reversed sentence pairs that contain the same components (Frankland & Greene 2015, experiment 1). Moreover, when sentences employing the same semantic components differ in their emotional salience (e.g., *the baby kicked the grandfather* versus *the grandfather kicked the baby*), the trial-by-trial strength of the response in the amygdala is mediated by variation in patterns of lmSTC activity, consistent with a role for lmSTC in sentence comprehension.

These findings dovetail those of Wu et al. (2007), who showed that damage to mid-left STS/STG causes deficits in understanding who did what to whom in response to both pictures and sentences (e.g., *the circle kissed the square*) (see also Dronkers et al. 2004 for relevant patient work). Moreover, Wang et al. (2016) found a region of the mid-STS in which patterns of activity enable classification of who did what to whom in a video. Although nearby regions of the left mid-anterior STG/STS have been implicated in linguistic functions pertaining to local aspects of sentence syntax (Friederici et al. 2003, Allen et al. 2012, Thothathiri & Rottinger 2015), the fact that such effects are observed using nonlinguistic stimuli (Wu et al. 2007, Wang et al. 2016) suggests that these representations may not be restricted to language. However, it is possible that people automatically convert to a linguistic representation to perform these tasks. Regardless of whether the representation is ultimately linguistic or nonlinguistic, critically, these studies do not provide evidence that the representation is compositional. The lmSTC might, for example, exhibit one random pattern for *the dog chased the cat* and another for *the cat chased the dog*, with no meaningful internal structure to either pattern.

In light of this, our second study (Frankland & Greene 2015, experiment 2) asked whether the lmSTC reuses representations in a manner consistent with compositional structure. Here, we used a pattern classifier to search within lmSTC for subregions encoding information about the identity of the agent (i.e., Who did it?). We performed a separate search for subregions encoding information about the patient (i.e., To whom was it done?). These analyses revealed two adjacent subregions: a medial subregion encoding information about the agent and a lateral subregion encoding information about the patient. Critically, the classifier was always tested on sentences with new verbs, implying that the representations for various agents and patients were reused across contexts. For example, in the patient region, the same pattern was observed for different meanings such as *the cat chased the dog* and *the dog was bumped by the boy*, where *dog* is the patient despite differences in both the linear order of the words and the other semantic elements involved. In subsequent work, we found that this distinction likely represents the underlying causal-temporal structure of the event rather than the syntactic structure of the sentence (subject/object) (Frankland & Greene 2019). This suggests the intriguing possibility that macroscopic topographic structure in the brain (from medial to lateral within this part of the temporal lobe) could reflect the "structural relations" of one aspect of the combinatorial structure envisaged by Fodor & Pylyshyn (1988). The idea of distinct neural populations representing agents and patients (or other semantic variables) fits well with a long research tradition in linguistics (e.g., Fillmore 1968, Jackendoff 1992; see Levin & Hovav 2005 for review) and with Marcus's (2001) argument that high-level cognition is algebraic, representing abstract variables that can be updated with new values to rapidly form new ideas.

We take the reuse of structured representations (noun-role bindings) in lmSTC across sentences to be a critical component of the composition of sentence meaning, as it both reflects structure (i.e., Who did it? To whom was it done?) and supports generalization to novel sentences. However, compositionality also requires such representational reuse one level down: It requires not only that the same concept-role bindings be reused, but, of course, that the same concepts occupy different roles, such that *dog* as agent bears some relation to *dog* as patient. We expect that these role-invariant representations are encoded elsewhere in the brain, perhaps in the posterior MTG or broadly distributed, and that the lmSTC represents compressed versions of these richer representations (see discussion of pointers in the section titled Separating Memory from Computation). However, how exactly the codes in the agent and patient regions relate to one another remains unclear (see also Frankland & Greene 2019) and is thus an important topic of future work.

The results described above involve bindings between concepts (e.g., *dog* or *cat*) and broad semantic roles (agent and patient). However, to fully comprehend an event, binding concepts to broad semantic roles is probably not enough. For example, in the more medial agent region, we saw no evidence that the encoding of *dog* in *the dog chased* differs from that of *dog* in *the dog scratched*. This is consistent with what the strictest version of the LoT hypothesis would predict, with neural patterns functioning like symbols that can be flexibly recombined. However, a dog that is chasing is different from dog that is scratching—for example, in its posture. If one is to imagine a dog chasing a cat, and answer certain questions about this event (e.g., Are the dog and the cat facing the same direction?), one must integrate the *dogness* and the *chasingness* in a plausible way. This suggests that elsewhere in the brain there should be representations that integrate representations of agents and actions, as well as actions and patients. This implies that categorical relations may be represented at different levels of abstraction.

In a more recent study (Frankland & Greene 2019), we tested this hypothesis by training separate voxel-wise encoding models to predict neural activity based on (*a*) broad semantic roles shared across verbs, as above (e.g., *dog-as-agent*) and (*b*) narrow semantic roles specific to a particular verb (e.g., *dog-as-chaser*). Here, we found a region of the rostral/frontal polar PFC (rPFC/BA10) that carries information about narrow noun-verb combinations, such as *dog-as-chaser* and *cat-as-chasee*. This is in contrast to the broad roles that structure lmSTC activity. This success of different encoding models in the lmSTC and rPFC may reflect a trade-off between abstraction (lmSTC) and specificity (rPFC) in event representation. Broad roles (represented in the lmSTC) support generalization to novel verbs and the mapping of event structure to sentence syntax (i.e., Who did it? To whom was it done?). Narrow roles (represented in the rPFC), by contrast, may provide the structured representations necessary to imagine and reason about specific events in a particular context. One might think of such representations as being like clip art (Frankland & Greene 2019): integrated semantic units that can be mixed and matched like symbols, but that also carry some information about how relationships are instantiated in the world. Critically, both lmSTC and rPFC reuse these entity-role bindings across sentences, supporting generalization to novel sentences.

The particular portion of rPFC that we identify is largely left lateralized and lies along the medial frontal gyrus (BA10). As with the lmSTC, this region is adjacent, but distinct, from the mPFC regions we discussed above in the DMN. Gilbert et al. (2010) note distinct contributions and functional connectivity for the lateral and medial portions of the rPFC. Lateral rPFC coactivates with the fronto-parietal control network, while medial rPFC coactivates with the DMN. The fronto-parietal control network is known to support cognitive and behavioral flexibility and fluid intelligence (Fuster 1997, Miller & Cohen 2001, Sigala et al. 2008, Duncan et al. 2017); by contrast, the DMN, as reviewed above, supports simulation sensitive to context (Schacter et al. 2008). Notably, the rPFC/BA10 region we identified as encoding narrow noun-verb combinations encompasses regions strongly connected to the DMN as well as regions strongly connected to fronto-parietal control regions, lying intermediate to the two. What is more, we see a similar connectivity pattern in the lmSTC, the region implicated in broad semantic role binding (e.g., *dog-as-agent*), which spans STG, STS, and MTG. Here, too, this patch of cortex spans regions that are more strongly tied to fronto-parietal control operations (STG) as well as regions tied to default mode function (MTG) (Pascual et al. 2013). Of course, anatomical positioning does not guarantee any functional relationship, or even the synaptic connections necessary to support it. Nevertheless, these observations suggest the intriguing possibility that the rPFC and lmSTC serve as representational waypoints between the highly flexible representational capacity of the lateral PFC (Kriete et al. 2013, Rigotti et al. 2013) and the map-like representations in the DMN (Doeller et al. 2010, Constantinescu et al. 2016).

Taken together, the results described in this section suggest that an important class of combinatorial conceptual processes recruits cortical regions beyond the conventional DMN. However, here, we make the stronger suggestion that the representations in these regions can fruitfully be considered sentence-like rather than map-like. Critically, sentence-like representations are built through the assignment of values to discrete, non-continuous variables, such as *agent* and *patient*, and therefore have an "algebraic" structure (Marcus 2001). Such representations allow for flexible recombination, the encoding of categorical directional relationships (e.g., a > b), and inferences based on form (syntax), independent of meaning (semantics): If the *glorp* was *bipped* by the *glax*, then the *glax bipped* the *glorp* (Fodor 1975, Fodor & Pylyshyn 1988, Pinker 1997). Map-like representations, by contrast, represent semantic content using location within a space defined by continuous variables. Because proximity in the representational space (e.g., adjacency on a map) corresponds to proximity in the represented domain, graded relational information is implicitly encoded, enabling inferences that are not well supported by coarser sentence-like representations, which may abstract away from this subtle semantic information. Such global, map-like representations may be critical for certain forms of planning (Behrens et al. 2018, Bellmund et al. 2018) as well as interpolation, extrapolation, and analogical reasoning (Frankland et al. 2019; J.D. Cohen, unpublished manuscript). This sentence/map distinction may be thought of as closely related to the distinction between digital and analog representations. Broadly, one might think of the lmSTC and rPFC as encoding compact, "digital" compositions of object/events in particular relations, which can both generate and be generated by the richer, "analog" conceptual compositions on the canvas of the DMN. This canvas may rely on a grid-like code for the global representations of currently relevant conceptual spaces. Understanding how these two classes of representation cooperate in the human brain is, we think, a central challenge for cognitive and computational neuroscience.

TOWARD COMPOSITIONALITY IN COMPUTATIONAL MODELS

We have described recent work on compositional processes in cognitive neuroscience, highlighting broad connections between these emerging literatures and classic theory positing a language of thought. Though we believe this integration offers valuable new perspectives and avenues for future work, a proper understanding of how the human brain combines concepts to construct complex, structured thoughts will require richer and more detailed computational models, in addition to broad theoretical frameworks. Recently, computational neuroscientists and artificial intelligence researchers have made progress toward developing systems that capture key features of compositional cognition. Here, we highlight progress on three fronts that are especially promising in our search for the algorithms that enable the highest forms of human thought: (*a*) generative models, (*b*) neural network architectures that separate memory from dynamic computation, and (*c*) systems that translate between representational formats, such as between visual and verbal representations.

Generative Models

Lake et al.'s (2015) Bayesian program learning (BPL) model of one-shot concept learning, though not a neural model, is important for present purposes because of its representational strategy. The BPL algorithm yields a hierarchical generative model of handwritten characters, drawn from unfamiliar alphabets. In contrast to a view of concepts as simple feature detectors in recognition systems, generative models both identify and construct (i.e., generate) samples. These models have been familiar in vision research, sometimes known as analysis-by-synthesis models (Yuille & Kersten 2006). Notably, Lake et al.'s (2015) model can categorize and construct members of a novel category after observing only one instance of the category—one-shot learning. In BPL, the character types are represented as a structured composition of a primitive set of reusable elements. This enables the representation of a novel concept (or conceptual combination) as a structured assembly of familiar subparts. To assemble these combinations, the model encodes the conditional probabilities of relations between subpart sequences—that is, it exploits structured knowledge about which subparts are likely to co-occur and follow other subparts in a particular configuration. Finally, it separates type-level representations of characters (abstract knowledge about the categories, which we expect to be represented throughout cortex) from token representations that are embedded in a particular image frame. This confluence between conceptual input and background knowledge concerning interactions between concepts in context is analogous to our suggested role for the DMN (see also Marcus 2001 on the centrality of the type/token distinction in compositional representation.)

Generative models are not restricted to a Bayesian framework, however. Neural networks, such as auto-encoders and Boltzmann machines, learn to encode, compress, and regenerate input data. Variational auto-encoders (VAE) (Kingma & Welling 2013) are a notable hybrid probabilistic/neural network model that uses flexible encoder/decoder neural networks to learn to generate samples from a low-dimensional latent space. Recent modifications of the VAE, such as the B-VAE (Higgins et al. 2017a), have been able to extract factorized representations of the input, with each node in the network's latent space representing a particular factor (or concept) orthogonal to the other nodes. If the appropriate architecture were employed to utilize these factors as input, these may be useful representations for conceptual combination (see also Higgins et al. 2017b).

Whittington et al. (2018) recently found that a recurrent VAE, tasked with predicting the next location of an entity navigating a graph, naturally learned periodic grid-like representations in the latent space of the network. Cueva & Wei (2018) had previously reported similar effects for a recurrent network tasked with predicting (x, y) location during navigation of a 2D space (see also Banino et al. 2018). Thus, neural networks with simple predictive objectives naturally develop grid-like representations, similar to the representations observed in the human DMN (Doeller et al. 2010, Constantinescu et al. 2016). That this particular representation emerges from recurrent generative networks is consistent with the slow temporal integration of information in the DMN (Hasson et al. 2008, Honey et al. 2012), as well as with the updates to the context representation proposed by Smith et al. (2018).

Separating Memory and Computation

The aforementioned generative models provide a useful framework for learning conceptual representations, both in Bayesian models (Lake et al. 2015) and, potentially, in neural networks (Higgins et al. 2017a,b; Whittington et al. 2018). Moreover, when tasked with prediction over time, they extract low-dimensional metric representations of the space, like grid cells (Banino et al. 2018, Cueva & Wei 2018, Whittington et al. 2018). We have suggested, however, that these grid-like representations may not be sufficient for representing explicit relations involving variable binding. How might the brain flexibly bind different pieces of information to abstract variables?

Some recent neural network models approximate the low-level operations of a silicon computer (Eliasmith 2013; Kriete et al. 2013; Graves et al. 2014, 2016; Hudson & Manning 2018), in some cases using biologically plausible models (Eliasmith 2013, Kriete et al. 2013, Crawford et al. 2016). A critical innovation of these models is to separate the dynamic aspects of control from stable memory representations, as in a classic Turing machine (Gallistel & King 2011). Mimicking symbolic architectures in this way has the potential to approximate variable binding in neural networks (Kriete et al. 2013, Graves et al. 2016).

One such approach is to employ pointers. Rather than operating on the entirety of a conceptual representation, in a pointer-based model, high-level cognitive processes operate over an address-like representation, which points to the location of that content in memory. The strategy of employing such pointers is known as "indirection" in computer science. In some models, this pointer is the location of the full representation (Kriete et al. 2013, Graves et al. 2014), but in others, the pointer representation carries compressed semantic content of what it points to (Eliasmith 2013, Graves et al. 2016). This distinction roughly corresponds to the distinction between location-based and content-based addressing in computers.

Kriete et al. (2013) developed a pointer-based model of variable binding based on the connectivity between anatomically separate stripes (which can be used for separate roles) in the DLPFC and the basal ganglia. Notably, using only a combination of reinforcement learning and error-driven learning, the network acquired the ability to construct novel role-filler combinations (e.g., encoding *John* as agent having only experienced *John* as patient). The same concept can thus be employed in different roles, with a pointer to the content housed in role-specific slots in the PFC. While this is a model of PFC function, it is possible that a similar architecture could support role-filler combinations in the agent/patient separation observed in the superior temporal cortex (Frankland & Greene 2015).

Similar abstract computational principles are at work in the recently developed differentiable neural computer (DNC) (Graves et al. 2016), the successor to the Neural Turing Machine (Graves et al. 2014). Unlike feed-forward neural networks, or even recurrent neural networks, DNCs reintroduce the Turing machine's separation between memory and computation in a trainable neural network. Moreover, unlike Turing machines (both classic and neural), the memory in a DNC is content addressable (Graves et al. 2016), and the weights are fully trainable using backpropagation. The network uses a separate neural network called a controller that receives and emits vectors over time, allowing it to interact with the memory matrix through read-and-write operations. Broadly, one can think of the vectors emitted by the controller as functioning like pointers, and portions of the memory representation as content. Or, put another way, one can think of the DNC's architecture as imposing a separation between operations (the actions of the controller) and operands (the content being processed). In this way, the DNC is able to accomplish a number of difficult structured tasks, such as copying novel sequences of unfamiliar length (Graves et al. 2014) and traversing a graph of the London Underground (Graves et al. 2016). Similar architectures can be used to separate operations from suboperations, dynamically routing information through distinct subnetworks to flexibly meet task demands (Rosenbaum et al. 2017, Cases et al. 2019).

Cross-Modal Translation of Compositional Representations

We have noted that the LoT hypothesis posits an amodal representational language, accessible to, but distinct from, both linguistic and perceptual input modalities. Here, we highlight two strands of recent neural network research that bear on the acquisition of such abstract, amodal representations: (*a*) attempts to perform visual question answering using the CLEVR dataset (Johnson et al. 2017) and (*b*) bidirectional generative models for producing linguistic descriptions from images, and vice versa.

CLEVR is a programmatically generated set of images displaying objects in complex spatial relations, along with sets of corresponding linguistic queries and answers. An example CLEVR question might be, "What size is the square that is left of the green metal thing that is right of the cylinder"? The questions require encoding multiple relations and are thus meant to capture complex visual reasoning. Remarkably, several neural network architectures have succeeded at this task, answering 97–99.5% of the questions correctly (Santoro et al. 2017, Hudson & Manning 2018, Yi et al. 2018). Santoro et al.'s (2017) model integrates the outputs of convolutional neural networks (specialized for object recognition) and a recurrent neural network [long short-term memory (LSTM)] devoted to processing the linguistic input using a relational decoder (see also Perez et al. 2018, Yi et al. 2018). The model thus comes equipped with architectural biases for processing linguistic and visual inputs, but the representations in all parts of the network are learned so as to minimize error on the relational reasoning task. To train the network end-to-end, the models may be provided the program that generated the image as a supervisory signal (Yi et al. 2018) or explicit category labels regarding the correct answer (e.g., *large*) (Santoro et al. 2017, Hudson & Manning 2018). Given that it is unlikely that humans have access to this type of information (and certainly not in this volume), it is worth considering neural network architectures that link language and perception without explicit supervision.

Bidirectional generative networks attempt to generate fragments of language from images (as in image captioning) and images from fragments of language (as in mental imagery). Many of these models minimize translation between perceptual input and linguistic output (and vice versa) by mapping in and out of a shared representational space, analogous to an amodal LoT (Srivastava & Salakhutdinov 2012, Kiros et al. 2014, Fang et al. 2015, Karpathy & Fei-Fei 2015, Mansimov et al. 2015). One notable case of image generation is provided by Mansimov et al. (2015). Here, the authors developed a bidirectional generative network that was capable of producing fairly plausible images corresponding to novel scene descriptions. For example, when asked to generate *a herd of elephants flying in blue skies*, the network generated samples of large gray objects over a blue background, despite having never encountered that particular composition. The strategy of mapping into a shared representational space is also used by translation models, such as English to French (Sutskever et al. 2014). However, unlike in a Fodorian LoT, the representational space is unconstrained, allowing backpropagation to freely acquire whatever representational states minimize reconstruction error. For engineering purposes, this is a virtue. However, proponents of the LoT hypothesis suspect that human comprehension depends on complex semantic representations with internal representations that are far more structurally constrained. LeCun et al. (2015, p. 441), however, disagree, writing that the success of deep learning on translation tasks "raises serious doubts about whether understanding a sentence requires anything like the internal symbolic expressions that are manipulated by using inference rules." Though we share their general excitement and optimism, we suggest that biological approximations of variable-binding and routing operations are better suited to model aspects of high-level human cognition (Eliasmith 2013, Kriete et al. 2013, Rosenbaum et al. 2017, Cases et al. 2019). For example, in recent modeling work, both Whittington et al. (2018) and Russin et al. (2019) have discovered direct benefits for compositional generalization by explicitly separating syntactic and semantic representations in recurrent networks, consistent with the LoT hypothesis. An important direction for future research is to systematically compare the representations acquired by neural networks in compositional domains to those employed by the human brain (see also Lake et al. 2017 for a thorough discussion of the limits of recent deep learning algorithms and proposals for moving forward).

SUMMARY AND CONCLUSION

The human brain exhibits a remarkable capacity to generate and comprehend complex ideas, including ideas that are new to us as individuals and new to the world entirely. This ability to make "infinite use of finite means" (Von Humboldt, cited in Chomsky 2006, p. 15) arises from the compositional nature of thought, whereby complex meanings are built out of simpler ones. In our attempt to understand how the brain builds complex meaning, we have revisited Fodor's (1975) LoT hypothesis, focusing on two of its key tenets. The first is the idea of a central, abstract representational system, separable from natural language and perception, that is shared across these different input modalities. The second is the stronger claim that this system is, in certain critical respects, structured like a language, employing combinatorial procedures that are distinct from the contents over which they operate, mirroring the distinction within natural language between syntax and semantics.

A broad range of evidence indicates that DMN-like regions play a critical role in conceptual combination at multiple levels, from representing the meanings of simple word pairs (Baron et al. 2010; Graves et al. 2010; Baron & Osherson 2011; Bemis & Pylkkänen 2011, 2013a; Pylkkänen et al. 2014; Price et al. 2015, 2016) to interpreting complex narratives (Honey et al. 2012, Regev et al. 2013, Tamir et al. 2015). The DMN enables the simulation of nonpresent episodes—past, future, or hypothetical (Buckner & Carroll 2007, Buckner et al. 2008, Schacter et al. 2008, De Brigard et al. 2013)—but its role in meaning making appears to be broader (Binder et al. 2009). Research on the timing of semantic processing indicates that the DMN is where semantic production begins and semantic comprehension ends (Bemis & Pylkkänen 2011, 2013a; Pylkkänen et al. 2014). Critically, for the LoT hypothesis, the DMN's representational format seems to be amodal, or multimodal, and to cut across conceptual domains (Fairhall & Caramazza 2013, Crittenden et al. 2015, Smith et al. 2018). Recent work suggests it represents implicit relational knowledge using cognitive maps (Behrens et al. 2018, Bellmund et al. 2018) that may depend on grid-like encodings for context-dependent updating of conceptual states. Such mechanisms, when applied to conceptual spaces, seem naturally suited to the representation of conceptual combinations and to the kinds of inferences—interpolations and extrapolations—that make mental simulations possible.

While the DMN comports well with the LoT's first tenet, it shows little sign of having a language-like representational structure. However, structures within the temporal lobe and prefrontal cortex show greater promise in realizing the LoT's second tenet: a system for flexibly representing explicitly structured conceptual combinations. Here, representations of who did what to whom follow the logic of language, binding representations of nouns to broad semantic roles (Wu et al. 2007; Frankland & Greene 2015, 2019; Wang et al. 2016). The building of such representations may be supported by the fronto-parietal control network, which approximates the sequential functionality of classical computer programs (Eliasmith 2013, Kriete et al. 2013, Graves et al. 2014). Thus, language-like representations in the temporal and frontal cortex may provide compact, abstract instructions for painting richer, more concrete compositions on the canvas of the DMN.

Across the topics surveyed here, theoretical inferences from human neuroscientific data remain largely speculative. However, the rapid advance of computational models of high-level, multimodal cognition promises to bring more stringent tests and new ideas to this emerging field. Generative models for concept learning (Lake et al. 2015), neural network models that separate computation from memory (Kriete et al. 2013, Graves et al. 2014, Hudson & Manning 2018), and hybrid neural networks for visual question answering and cross-modal translation (Srivastava & Salakhutdinov 2012, Kiros et al. 2014, Mansimov et al. 2015, Santoro et al. 2017) are providing clues as to how the neurons in our own heads might implement a language of thought.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Dan Schacter for helpful comments.

LITERATURE CITED

Addis DR, Wong AT, Schacter DL. 2007. Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologia* 45(7):1363–77

Allen K, Pereira F, Botvinick M, Goldberg AE. 2012. Distinguishing grammatical constructions with fMRI pattern analysis. *Brain Lang.* 123(3):174–82

Amodio DM, Frith CD. 2006. Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* 7:268–77

Anderson JR. 1978. Arguments concerning representations for mental imagery. *Psychol. Rev.* 85(4):249–77

Andrews-Hanna JR, Reidler JS, Sepulcre J, Poulin R, Buckner RL. 2010. Functional-anatomic fractionation of the brain's default network. *Neuron* 65(4):550–62

Aronov D, Nevers R, Tank DW. 2017. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature* 543(7647):719–22

Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. 2017. Discovering event structure in continuous narrative perception and memory. *Neuron* 95(3):709–21

Banino A, Barry C, Uria B, Blundell C, Lillicrap T, et al. 2018. Vector-based navigation using grid-like representations in artificial agents. *Nature* 557(7705):429–33

Baron SG, Osherson D. 2011. Evidence for conceptual combination in the left anterior temporal lobe. *Neuroimage* 55(4):1847–52

Baron SG, Thompson-Schill SL, Weber M, Osherson D. 2010. An early stage of conceptual combination: superimposition of constituent concepts in left anterolateral temporal lobe. *Cogn. Neurosci.* 1(1):44–51

Barron HC, Dolan RJ, Behrens TE. 2013. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat. Neurosci.* 16:1492–98

Barry C, Hayman R, Burgess N, Jeffery KJ. 2007. Experience-dependent rescaling of entorhinal grids. *Nat. Neurosci.* 10(6):682

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):11347–53

Behrens TE, Muller TH, Whittington JC, Mark S, Baram AB, et al. 2018. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* 100(2):490–509

Bellmund JL, Deuker L, Schröder TN, Doeller CF. 2016. Grid-cell representations in mental simulation. *eLife* 5:e17089

Bellmund JL, Gärdenfors P, Moser EI, Doeller CF. 2018. Navigating cognition: spatial codes for human thinking. *Science* 362(6415):eaat6766

Bemis DK, Pylkkänen L. 2011. Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J. Neurosci.* 31(8):2801–14

Bemis DK, Pylkkänen L. 2013a. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cereb. Cortex* 23(8):1859–73

Bemis DK, Pylkkänen L. 2013b. Combination across domains: an MEG investigation into the relationship between mathematical, pictorial, and linguistic processing. *Front. Psychol.* 3:583

Benoit RG, Schacter DL. 2015. Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia* 75:450–57

Binder JR, Desai RH, Graves WW, Conant LL. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19(12):2767–96

Boylan C, Trueswell JC, Thompson-Schill SL. 2015. Compositionality and the angular gyrus: a multi-voxel similarity analysis of the semantic composition of nouns and verbs. *Neuropsychologia* 78:130–41

Braga RM, Buckner RL. 2017. Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron* 95(2):457–71

Buckner RL, Andrews-Hanna JR, Schacter DL. 2008. The brain's default network. *Ann. N. Y. Acad. Sci.* 1124(1):1–38

Buckner RL, Carroll DC. 2007. Self-projection and the brain. *Trends Cogn. Sci.* 11(2):49–57

Buckner RL, Snyder AZ, Shannon BJ, LaRossa G, Sachs R, et al. 2005. Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *J. Neurosci.* 25:7709–17

Camp E. 2007. Thinking with maps. *Philos. Perspect.* 21(1):145–82

Carpenter F, Manson D, Jeffery K, Burgess N, Barry C. 2015. Grid cells form a global representation of connected environments. *Curr. Biol.* 25(9):1176–82

Cases I, Rosenbaum C, Riemer M, Geiger A, Klinger T, et al. 2019. Recursive routing networks: learning to compose modules for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3631–48. Stroudsburg, PA: ACL

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):115–25

Chomsky N. 2006. *Language and Mind*. Cambridge, UK: Cambridge Univ. Press

Christoff K, Gordon AM, Smallwood J, Smith R, Schooler JW. 2009. Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *PNAS* 106:8719–24

Christoff K, Irving ZC, Fox KC, Spreng RN, Andrews-Hanna JR. 2016. Mind-wandering as spontaneous thought: a dynamic framework. *Nat. Rev. Neurosci.* 17:718–31

Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, et al. 2012. The representation of biological classes in the human brain. *J. Neurosci.* 32(8):2608–18

Constantinescu AO, O'Reilly JX, Behrens TE. 2016. Organizing conceptual knowledge in humans with a gridlike code. *Science* 352(6292):1464–68

Coutanche MN, Thompson-Schill SL. 2014. Creating concepts from converging features in human cortex. *Cereb. Cortex* 25(9):2584–93

Crawford E, Gingerich M, Eliasmith C. 2016. Biologically plausible, human-scale knowledge representation. *Cogn. Sci.* 40(4):782–821

Crittenden BM, Mitchell DJ, Duncan J. 2015. Recruitment of the default mode network during a demanding act of executive control. *eLife* 4:e06481

Cueva CJ, Wei XX. 2018. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. arXiv:1803.07770 [q-bio.NC]

D'Argembeau A, Collette F, Van der Linden M, Laureys S, Del Fiore G, et al. 2005. Self-referential reflective activity and its relationship with rest: a PET study. *Neuroimage* 25(2):616–24

De Brigard F, Addis DR, Ford JH, Schacter DL, Giovanello KS. 2013. Remembering what could have happened: neural correlates of episodic counterfactual thinking. *Neuropsychologia* 51(12):2401–14

Del Prato P, Pylkkänen L. 2014. MEG evidence for conceptual combination but not numeral quantification in the left anterior temporal lobe during language production. *Front. Psychol.* 5:524

Devereux BJ, Clarke A, Marouchos A, Tyler LK. 2013. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *J. Neurosci.* 33(48):18906–16

Doeller CF, Barry C, Burgess N. 2010. Evidence for grid cells in a human memory network. *Nature* 463(7281):657–61

Dordek Y, Soudry D, Meir R, Derdikman D. 2016. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife* 5:e10094

Doumas LA, Hummel JE. 2012. Computational models of higher cognition. In *The Oxford Handbook of Thinking and Reasoning*, ed. KJ Holyoak, RG Morrison, pp. 52–66. Oxford, UK: Oxford Univ. Press

Doumas LA, Hummel JE, Sandhofer CM. 2008. A theory of the discovery and predication of relational concepts. *Psychol. Rev.* 115(1):1–43

Dronkers NF, Wilkins DP, Van Valin RD Jr., Redfern BB, Jaeger JJ. 2004. Lesion analysis of the brain areas involved in language comprehension. *Cognition* 92(1–2):145–77

Duncan J, Chylinski D, Mitchell DJ, Bhandari A. 2017. Complexity and compositionality in fluid intelligence. *PNAS* 114(20):5295–99

Eliasmith C. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford, UK: Oxford Univ. Press

Estes Z, Glucksberg S. 2000. Interactive property attribution in concept combination. *Mem. Cogn.* 28(1):28–34

Evans G. 1982. *The Varieties of Reference*. Oxford, UK: Oxford Univ. Press

Fairhall SL, Caramazza A. 2013. Brain regions that represent amodal conceptual knowledge. *J. Neurosci.* 33(25):10552–58

Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473–82. New York: IEEE

Fiete IR, Burak Y, Brookings T. 2008. What grid cells convey about rat location. *J. Neurosci.* 28(27):6858–71

Fillmore CJ. 1968. The case for case. In *Universals in Linguistic Theory*, ed. E Bach, RT Harms, pp. 1–88. New York: Holt, Rinehart, and Winston

Fodor JA. 1975. *The Language of Thought*, Vol. 5. Cambridge, MA: Harvard Univ. Press

Fodor JA. 2001. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press

Fodor JA. 2003. *Hume Variations*. Oxford, UK: Oxford Univ. Press

Fodor JA, Pylyshyn ZW. 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition* 28(1–2):3–71

Forgács B, Bohrn I, Baudewig J, Hofmann MJ, Pléh C, Jacobs AM. 2012. Neural correlates of combinatorial semantic processing of literal and figurative noun noun compound words. *Neuroimage* 63(3):1432–42

Frankland SM, Greene JD. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *PNAS* 112(37):11732–37

Frankland SM, Greene JD. 2019. Two ways to build a thought: distinct forms of compositional semantic representation across brain regions. PsyArXiv, Aug. 27. **https://doi.org/10.31234/osf.io/65tn7**

Frankland SM, Webb TW, Petrov AA, O'Reilly RC, Cohen JD. 2019. Extracting and utilizing abstract, structured representations for analogy. In *Proceedings of the Cognitive Science Society*, pp. 1766–72. Montreal: Cogn. Sci. Soc.

Frege G. 1980. Letter to Jourdain. In *Philosophical and Mathematical Correspondence of Gottlob Frege*, ed. B McGuinness, pp. 78–80. Chicago: Univ. Chicago Press

Friederici AD, Rueschemeyer SA, Hahne A, Fiebach CJ. 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cereb. Cortex* 13(2):170–77

Fuster J. 1997. *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*. New York: Lippincott-Raven

Gagné CL, Shoben EJ. 1997. Influence of thematic relations on the comprehension of modifier–noun combinations. *J. Exp. Psychol. Learn. Mem. Cogn.* 23(1):71–87

Gallistel CR, King AP. 2011. *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*, Vol. 6. Hoboken, NJ: John Wiley & Sons

Gärdenfors P. 2004. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press

Garner HL. 1959. The residue number system. In *Papers Presented at the March 3–5, 1959, Western Joint Computer Conference*, pp. 146–53. New York: ACM

Garvert MM, Dolan RJ, Behrens TE. 2017. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife* 6:e17086

Gilbert SJ, Gonen-Yaacovi G, Benoit RG, Volle E, Burgess PW. 2010. Distinct functional connectivity associated with lateral versus medial rostral prefrontal cortex: a meta-analysis. *Neuroimage* 53(4):1359–67

Goodman ND, Tenenbaum JB, Gerstenberg T. 2014. *Concepts in a Probabilistic Language of Thought*. Cambridge, MA: CBMM Memos

Graves A, Wayne G, Danihelka I. 2014. Neural Turing machines. arXiv:1410.5401 [cs.NE]

Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature* 538:471–76

Graves WW, Binder JR, Desai RH, Conant LL, Seidenberg MS. 2010. Neural correlates of implicit and explicit combinatorial semantic processing. *Neuroimage* 53(2):638–46

Gu S, Pasqualetti F, Cieslak M, Telesford QK, Alfred BY, et al. 2015. Controllability of structural brain networks. *Nat. Commun.* 6:8414

Hafting T, Fyhn M, Molden S, Moser MB, Moser EI. 2005. Microstructure of a spatial map in the entorhinal cortex. *Nature* 436(7052):801–6

Hagoort P, Indefrey P. 2014. The neurobiology of language beyond single words. *Annu. Rev. Neurosci.* 37:347–62

Hampton JA. 1997. Conceptual combination. In *Knowledge, Concepts and Categories*, ed. K Lambert, D Shanks, pp. 133–59. Sussex, UK: Psychol. Press

Hassabis D, Kumaran D, Maguire EA. 2007. Using imagination to understand the neural basis of episodic memory. *J. Neurosci.* 27(52):14365–74

Hassabis D, Maguire EA. 2007. Deconstructing episodic memory with construction. *Trends Cogn. Sci.* 11(7):299–306

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. 2008. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* 28(10):2539–50

Hayworth KJ. 2012. Dynamically partitionable autoassociative networks as a solution to the neural binding problem. *Front. Comput. Neurosci.* 6:73

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, et al. 2017a. β-*VAE: learning basic visual concepts with a constrained variational framework*. Paper presented at the 5th International Conference on Learning Representations, Toulon, Fr.

Higgins I, Sonnerat N, Matthey L, Pal A, Burgess C, et al. 2017b. Scan: learning abstract hierarchical compositional visual concepts. arXiv:1707.03389 [stat.ML]

Honey CJ, Thesen T, Donner TH, Silbert LJ, Carlson CE, et al. 2012. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* 76(2):423–34

Horner AJ, Bisby JA, Zotow E, Bush D, Burgess N. 2016. Grid-like processing of imagined navigation. *Curr. Biol.* 26:842–47

Hudson DA, Manning CD. 2018. Compositional attention networks for machine reasoning. arXiv:1803.03067 [cs.AI]

Hume D. 2003 (1738). *A Treatise of Human Nature*. North Chelmsford, MA: Courier Corp.

Hummel JE, Holyoak KJ. 1997. Distributed representations of structure: a theory of analogical access and mapping. *Psychol. Rev.* 104(3):427–66

Hummel JE, Holyoak KJ, Green C, Doumas LA, Devnich D, et al. 2004. A solution to the binding problem for compositional connectionism. In *Compositional Connectionism in Cognitive Science: Papers from the AAAI Fall Symposium*, ed. SD Levy, R Gayler, pp. 31–34. Palo Alto, CA: AAAI

Humphreys GF, Hoffman P, Visser M, Binney RJ, Ralph MAL. 2015. Establishing task-and modality-dependent dissociations between the semantic and default mode networks. *PNAS* 112(25):7857–62

Humphreys GF, Ralph MAL. 2014. Fusion and fission of cognitive functions in the human parietal cortex. *Cereb. Cortex* 25(10):3547–60

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532(7600):453–58

Irish M, Addis DR, Hodges JR, Piguet O. 2012. Considering the role of semantic memory in episodic future thinking: evidence from semantic dementia. *Brain* 135(7):2178–91

Irish M, Piguet O. 2013. The pivotal role of semantic memory in remembering the past and imagining the future. *Front. Behav. Neurosci.* 7:27

Jackendoff R. 1992. *Semantic Structures*. Cambridge, MA: MIT Press

Jacobs J, Weidemann CT, Miller JF, Solway A, Burke JF, et al. 2013. Direct recordings of grid-like neuronal activity in human spatial navigation. *Nat. Neurosci.* 16(9):1188–90

Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Lawrence Zitnick C, Girshick R. 2017. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–10. New York: IEEE

Johnson K. 2004. On the systematicity of language and thought. *J. Philos.* 101(3):111–39

Johnson SC, Baxter LC, Wilder LS, Pipe JG, Heiserman JE, Prigatano GP. 2002. Neural correlates of self-reflection. *Brain* 125(8):1808–14

Karpathy A, Fei-Fei L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–37. New York: IEEE

Kingma DP, Welling M. 2013. Auto-encoding variational Bayes. arXiv:1312.6114 [stat.ML]

Kiros R, Salakhutdinov R, Zemel R. 2014. Multimodal neural language models. *PMLR* 32(2):595–603

Konkle T, Caramazza A. 2013. Tripartite organization of the ventral stream by animacy and object size. *J. Neurosci.* 33(25):10235–42

Kriete T, Noelle DC, Cohen JD, O'Reilly RC. 2013. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *PNAS* 110(41):16390–95

Lake BM, Salakhutdinov R, Tenenbaum JB. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–38

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2017. Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44

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):2906–15

Leshinskaya A, Caramazza A. 2015. Abstract categories of functions in anterior parietal lobe. *Neuropsychologia* 76:27–40

Levin B, Hovav MR. 2005. *Argument Realization*. Cambridge, UK: Cambridge Univ. Press

Maguire EA, Burgess N, Donnett JG, Frackowiak RS, Frith CD, O'Keefe J. 1998. Knowing where and getting there: a human navigation network. *Science* 280:921–24

Mahon BZ, Anzellotti S, Schwarzbach J, Zampini M, Caramazza A. 2009. Category-specific organization in the human brain does not require visual experience. *Neuron* 63(3):397–405

Mahon BZ, Caramazza A. 2009. Concepts and categories: a cognitive neuropsychological perspective. *Annu. Rev. Psychol.* 60:27–51

Mansimov E, Parisotto E, Ba JL, Salakhutdinov R. 2015. Generating images from captions with attention. arXiv:1511.02793 [cs.LG]

Marcus GF. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press

Margulies DS, Ghosh SS, Goulas A, Falkiewicz M, Huntenburg JM, et al. 2016. Situating the default-mode network along a principal gradient of macroscale cortical organization. *PNAS* 113(44):12574–79

Martin A. 2007. The representation of object concepts in the brain. *Annu. Rev. Psychol.* 58:25–45

Martin A, Chao LL. 2001. Semantic memory and the brain: structure and processes. *Curr. Opin. Neurobiol.* 11(2):194–201

Mazoyer B, Zago L, Mellet E, Bricogne S, Etard O, et al. 2001. Cortical networks for working memory and executive functions sustain the conscious resting state in man. *Brain Res. Bull.* 54(3):287–98

Medin DL, Shoben EJ. 1988. Context and structure in conceptual combination. *Cogn. Psychol.* 20(2):158–90

Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24(1):167–202

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191–95

Molinaro N, Carreiras M, Duñabeitia JA. 2012. Semantic combinatorial processing of non-anomalous expressions. *Neuroimage* 59(4):3488–501

Molinaro N, Paz-Alonso PM, Duñabeitia JA, Carreiras M. 2015. Combinatorial semantics strengthens angular-anterior temporal coupling. *Cortex* 65:113–27

Murphy GL. 1988. Comprehending complex concepts. *Cogn. Sci.* 12(4):529–62

Musslick S, Saxe A, Özcimder K, Dey B, Henselman G, Cohen JD. 2017. Multitasking capability versus learning efficiency in neural network architectures. In *Proceedings of the Cognitive Science Society*, pp. 829–34. Austin, TX: Cogn. Sci. Soc.

Newell A. 1980. Physical symbol systems. *Cogn. Sci.* 4(2):135–83

O'Reilly RC, Busby RS. 2002. Generalizable relational binding from coarse-coded distributed representations. In *Advances in Neural Information Processing Systems 14*, ed. DG Dietterich, S Becker, Z Ghahramani, pp. 75–82. San Diego, CA: NIPS

Pallier C, Devauchelle AD, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. *PNAS* 108(6):2522–27

Pascual B, Masdeu JC, Hollenbeck M, Makris N, Insausti R, et al. 2013. Large-scale brain networks of the human left temporal pole: a functional connectivity MRI study. *Cereb. Cortex* 25(3):680–702

Patterson K, Nestor PJ, Rogers TT. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* 8(12):976–87

Pearson J, Kosslyn SM. 2015. The heterogeneity of mental representation: ending the imagery debate. *PNAS* 112(33):10089–92

Penn DC, Holyoak KJ, Povinelli DJ. 2008. Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* 31(2):109–30

Perez E, Strub F, De Vries H, Dumoulin V, Courville A. 2018. Film: visual reasoning with a general conditioning layer. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3942–51. Palo Alto, CA: AAAI

Piantadosi ST. 2011. *Learning and the Language of Thought*. PhD Diss., Mass. Inst. Technol., Cambridge, MA

Piantadosi ST, Jacobs RA. 2016. Four problems solved by the probabilistic language of thought. *Curr. Dir. Psychol. Sci.* 25(1):54–59

Pinker S. 1997. *How the Mind Works*. New York: W.W Norton & Co.

Pinker S. 2005. So how does the mind work? *Mind Lang*. 20(1):1–24

Plate TA. 1995. Holographic reduced representations. *IEEE Trans. Neural Netw.* 6(3):623–41

Poortman EB, Pylkkänen L. 2016. Adjective conjunction as a window into the LATL's contribution to conceptual combination. *Brain Lang*. 160:50–60

Price AR, Bonner MF, Peelle JE, Grossman M. 2015. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *J. Neurosci.* 35(7):3276–84

Price AR, Peelle JE, Bonner MF, Grossman M, Hamilton RH. 2016. Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. *J. Neurosci.* 36(13):3829–38

Pylkkänen L, Bemis DK, Elorrieta EB. 2014. Building phrases in language production: an MEG study of simple composition. *Cognition* 133(2):371–84

Rabagliati H, Doumas LA, Bemis DK. 2017. Representing composed meanings through temporal binding. *Cognition* 162:61–72

Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. 2001. A default mode of brain function. *PNAS* 98(2):676–82

Ralph MAL, Jefferies E, Patterson K, Rogers TT. 2017. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* 18(1):42–55

Regev M, Honey CJ, Simony E, Hasson U. 2013. Selective and invariant neural responses to spoken and written narratives. *J. Neurosci.* 33(40):15978–88

Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, et al. 2013. The importance of mixed selectivity in complex cognitive tasks. *Nature* 497(7451):585–90

Rosenbaum C, Klinger T, Riemer M. 2017. Routing networks: adaptive selection of non-linear functions for multi-task learning. arXiv:1711.01239 [cs.LG]

Rougier NP, Noelle DC, Braver TS, Cohen JD, O'Reilly RC. 2005. Prefrontal cortex and flexible cognitive control: rules without symbols. *PNAS* 102(20):7338–43

Russin J, Jo J, O'Reilly RC. 2019. Compositional generalization in a deep seq2seq model by separating syntax and semantics. arXiv:1904.09708 [cs.LG]

Santoro A, Raposo D, Barrett DG, Malinowski M, Pascanu R, et al. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30*, ed. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, et al., pp. 4967–76. San Diego, CA: NIPS

Saxe R, Carey S, Kanwisher N. 2004. Understanding other minds: linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.* 55:87–124

Schacter DL, Addis DR, Buckner RL. 2007. Remembering the past to imagine the future: the prospective brain. *Nat. Rev. Neurosci.* 8(9):657–61

Schacter DL, Addis DR, Buckner RL. 2008. Episodic simulation of future events: concepts, data, and applications. *Ann. N. Y. Acad. Sci.* 1124(1):39–60

Schacter DL, Addis DR, Hassabis D, Martin VC, Spreng RN, Szpunar KK. 2012. The future of memory: remembering, imagining, and the brain. *Neuron* 76(4):677–94

Shadlen MN, Movshon JA. 1999. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* 24(1):67–77

Shastri L, Ajjanagadde V. 1993. From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behav. Brain Sci.* 16(3):417–51

Shepard RN. 1987. Toward a universal law of generalization for psychological science. *Science* 237(4820):1317–23

Shulman GL, Fiez JA, Corbetta M, Buckner RL, Miezin FM, et al. 1997. Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. *J. Cogn. Neurosci.* 9(5):648–63

Sigala N, Kusunoki M, Nimmo-Smith I, Gaffan D, Duncan J. 2008. Hierarchical coding for sequential task events in the monkey prefrontal cortex. *PNAS* 105(33):11969–74

Smith EE, Osherson DN. 1984. Conceptual combination with prototype concepts. *Cogn. Sci.* 8(4):337–61

Smith EE, Osherson DN, Rips LJ, Keane M. 1988. Combining prototypes: a selective modification model. *Cogn. Sci.* 12(4):485–527

Smith V, Mitchell DJ, Duncan J. 2018. Role of the default mode network in cognitive transitions. *Cereb. Cortex* 28(10):3685–96

Smolensky P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.* 46(1–2):159–216

Spreng RN, Stevens WD, Chamberlain JP, Gilmore AW, Schacter DL. 2010. Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *Neuroimage* 53:303–17

Srivastava N, Salakhutdinov RR. 2012. Multimodal learning with deep Boltzmann machines. In *Advances in Neural Information Processing Systems 25*, ed. F Pereira, CJC Burges, L Bottou, KQ Weinberger, pp. 2222–30. San Diego, CA: NIPS

Stachenfeld KL, Botvinick MM, Gershman SJ. 2017. The hippocampus as a predictive map. *Nat. Neurosci.* 20(11):1643–53

Sutskever I, Vinyals O, Le QV. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 3104–12. San Diego, CA: NIPS

Tamir DI, Bricker AB, Dodell-Feder D, Mitchell JP. 2015. Reading fiction and reading minds: the role of simulation in the default network. *Soc. Cogn. Affect. Neurosci.* 11(2):215–24

Thompson-Schill SL. 2003. Neuroimaging studies of semantic memory: inferring "how" from "where." *Neuropsychologia* 41(3):280–92

Thothathiri M, Rattinger M. 2015. Ventral and dorsal streams for choosing word order during sentence production. *PNAS* 112(50):15456–61

Tolman EC. 1948. Cognitive maps in rats and men. *Psychol. Rev.* 55(4):189–208

Tversky A. 1977. Features of similarity. *Psychol. Rev.* 84(4):327–52

Van der Velde F, De Kamps M. 2006. Neural blackboard architectures of combinatorial structures in cognition. *Behav. Brain Sci.* 29(1):37–70

Vincent JL, Snyder AZ, Fox MD, Shannon BJ, Andrews JR, et al. 2006. Coherent spontaneous activity identifies a hippocampal-parietal memory network. *J. Neurophysiol.* 96:3517–31

Wang J, Cherkassky VL, Yang Y, Chang KMK, Vargas R, et al. 2016. Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. *Cogn. Neuropsychol.* 33(3–4):257–64

Westerlund M, Kastner I, Al Kaabi M, Pylkkänen L. 2015. The LATL as locus of composition: MEG evidence from English and Arabic. *Brain Lang.* 141:124–34

Westerlund M, Pylkkänen L. 2014. The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia* 57:59–70

Whittington JC, Muller TH, Barry C, Behrens TE. 2018. Generalisation of structural knowledge in the hippocampal-entorhinal system. arXiv:1805.09042 [cs.AI]

Wisniewski EJ. 1997. When concepts combine. *Psychon. Bull. Rev.* 4(2):167–83

Wisniewski EJ, Middleton EL. 2002. Of bucket bowls and coffee cup bowls: spatial alignment in conceptual combination. *J. Mem. Lang.* 46(1):1–23

Wu DH, Waller S, Chatterjee A. 2007. The functional neuroanatomy of thematic role and locative relational knowledge. *J. Cogn. Neurosci.* 19(9):1542–55

Wurm MF, Caramazza A. 2019. Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. *Nat. Commun.* 10(1):289

Yi K, Wu J, Gan C, Torralba A, Kohli P, Tenenbaum J. 2018. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems 31*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 1039–50. San Diego, CA: NIPS

Yuille A, Kersten D. 2006. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* 10(7):301–8

Zadbood A, Chen J, Leong YC, Norman KA, Hasson U. 2017. How we transmit memories to other brains: constructing shared neural representations via communication. *Cereb. Cortex* 27(10):4988–5000

Zhang L, Pylkkänen L. 2015. The interplay of composition and concept specificity in the left anterior temporal lobe: an MEG study. *Neuroimage* 111:228–40

Ziegler J, Pylkkänen L. 2016. Scalar adjectives and the temporal unfolding of semantic composition: an MEG investigation. *Neuropsychologia* 89:161–71