

# The Relational Bottleneck as an Inductive Bias for Efficient Abstraction

Taylor W. Webb<sup>1,\*</sup>, Steven M. Frankland<sup>2</sup>, Awni Altabaa<sup>3</sup>, Kamesh Krishnamurthy<sup>4</sup>,  
Declan Campbell<sup>4</sup>, Jacob Russin<sup>5</sup>, Randall O'Reilly<sup>6</sup>, John Lafferty<sup>3</sup>, and Jonathan  
D. Cohen<sup>4</sup>

<sup>1</sup>University of California, Los Angeles

<sup>2</sup>Dartmouth College

<sup>3</sup>Yale University

<sup>4</sup>Princeton University

<sup>5</sup>Brown University

<sup>6</sup>University of California, Davis

\*Correspondence to: taylor.w.webb@gmail.com

## Abstract

A central challenge for cognitive science is to explain how abstract concepts are acquired from limited experience. This effort has often been framed in terms of a dichotomy between empiricist and nativist approaches, most recently embodied by debates concerning deep neural networks and symbolic cognitive models. Here, we highlight a recently emerging line of work that suggests a novel reconciliation of these approaches, by exploiting an inductive bias that we term the *relational bottleneck*. We review a family of models that employ this approach to induce abstractions in a data-efficient manner, emphasizing their potential as candidate models for the acquisition of abstract concepts in the human mind and brain.

## Highlights

- Human learners efficiently acquire abstract concepts from limited experience. The effort to explain this capacity has fueled debate between proponents of symbolic and connectionist approaches, and motivated proposals for hybrid neuro-symbolic systems.
- A recently emerging approach, that we term the ‘relational bottleneck’ principle, suggests a novel way to bridge the gap. We formulate this in information theoretic terms, and review neural network architectures that implement this principle, displaying rapid learning of relational patterns, and systematic generalization of those patterns to novel inputs.
- The approach may help to explain a diverse set of phenomena, ranging from cognitive development to capacity limits in cognitive function. The approach is also consistent with findings from cognitive neuroscience, and may offer a useful general principle for designing more powerful artificial learning systems.

## Modeling the efficient induction of abstractions

Human cognition is characterized by a remarkable ability to transcend the specifics of limited experience to entertain highly general, abstract ideas. Understanding how the mind and brain accomplish this has been a central challenge throughout the history of cognitive science, and a major preoccupation of philosophy before that [1, 2, 3, 4]. Of particular importance is the central role played by *relations*, which enable human reasoners to abstract away from the details of individual entities and identify higher-order patterns across distinct domains [5, 6]. The capacity to think in terms of relations is a major

component underlying the human capacity for fluid reasoning [7, 8], and a key factor distinguishing human intelligence from that of other species [9].

Efforts to explain this capacity have often been framed in terms of a debate between **empiricism** (see Glossary), according to which concepts are primarily acquired from experience, and **nativism**, according to which certain core concepts are innate. Cognitive scientists in the empiricist tradition have for decades explored how the abstractions associated with human cognition might emerge through experience in neural architectures using general-purpose learning algorithms (often termed **connectionism**) [10, 11, 12, 13]. This endeavor has recently taken on new relevance, as the success of large language models has demonstrated that it is possible, in some cases, for a human-like capacity for abstraction to emerge given sufficient scaling of both architecture and training data [14, 15, 16, 17]. For instance, it has recently been shown that large language models can solve various analogy problems at a level equal to that of college students [18]. However, the ability of these models to perform abstract tasks (e.g., analogy) depends on exposure to a much larger training corpus than individual humans receive in an entire lifetime [19, 20].

An alternative approach, often associated with the nativist tradition, holds that human cognition arises from processes akin to symbolic programs. This approach has a long tradition in both cognitive science and AI [21, 22, 23], due to the fact that it offers a natural explanation of the flexibility of human cognition: processes that operate over symbols can be sensitive to their general structure, without respect to a particular symbol’s reference [24]. Recent efforts have demonstrated that this approach is capable of inducing abstract concepts in a data-efficient manner, mirroring the efficiency of human concept learning [25, 26, 27, 28, 29, 30, 31]. However, a potential limitation of this approach is that it depends on the pre-specification of symbolic primitives. Though it remains to be seen how far this approach can be scaled, it has thus far proven challenging to identify a set of primitives that are expressive enough to account for the breadth and richness of human natural concepts. It also raises the question of how the symbolic primitives arise: are these an innate feature of the human mind, or could they too emerge in a data-efficient manner through learning?

In this review, we highlight a recently emerging approach that suggests a novel reconciliation of these two traditions. The central feature of this approach is an **inductive bias** that we refer to as the *relational bottleneck*: a constraint that biases neural network models to focus on relations between objects rather than the attributes of individual objects. This approach satisfies two key desiderata for models of abstract concept acquisition. First, the approach is capable of rapid acquisition of abstract relational concepts. Second, the approach does not require access to a set of pre-specified primitives. This latter feature distinguishes the approach from other so-called *neuro-symbolic* approaches (see Box 2 for further discussion). In the following sections, we first provide a general characterization of this approach, drawing on concepts from information theory, and discuss a number of recently proposed neural network architectures that implement the approach. We then discuss the potential of the approach for modeling human cognition, relating it to existing cognitive theories and considering potential mechanisms through which it might be implemented in the brain.

## The relational bottleneck

We define the relational bottleneck as any mechanism that restricts the flow of information from perceptual to downstream reasoning systems to consist only of relations (see Box 1 for a formal definition). For example, given inputs representing individual visual objects, a relational bottleneck would constrain the representations passed to downstream reasoning processes such that they capture only the relations between these objects (e.g., whether the objects have the same shape), rather than the individual features of the objects (e.g., the individual shapes). Such a representation encourages downstream processes to identify relational patterns, such as the identity rule in Figure 1, in a manner that is abstracted away from the details of specific instances of those patterns, and can therefore be systematically generalized to novel inputs. In the following section, we highlight three recently proposed neural network architectures that instantiate this approach in different guises, illustrating how they utilize a relational bottleneck to induce abstract concepts in a data-efficient manner.

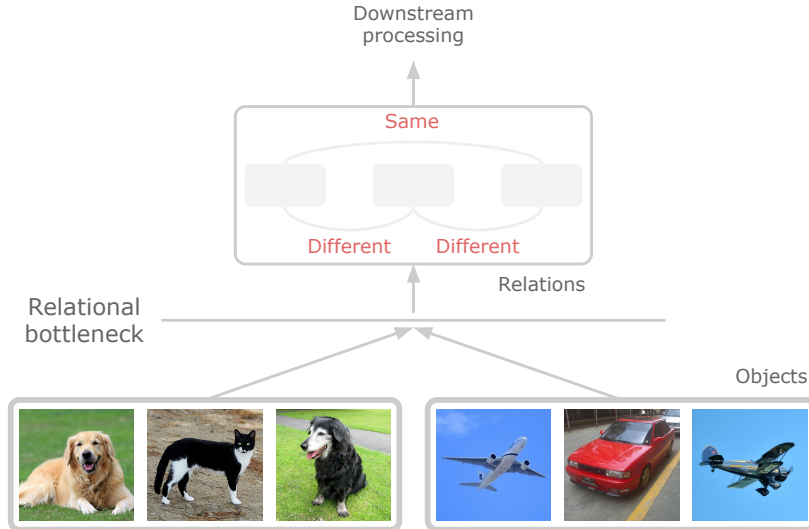


Figure 1: **The relational bottleneck.** An inductive bias that prioritizes the representation of relations (e.g., ‘same’ vs. ‘different’), and discourages the representation of the features of individual objects (e.g., the shape or color of the objects in the images above). The result is that downstream processing is driven primarily, or even exclusively by patterns of relations, and can therefore systematically generalize those patterns across distinct instances (e.g., the common ABA pattern displayed on both left and right), even for completely novel objects. The approach is illustrated here with same/different relations, but other relations can also be accommodated. Note that this example is intended only to illustrate the overall goal of the relational bottleneck framework. Figure 2 depicts neural architectures that implement the approach.

---

### Box 1: The relational bottleneck principle

Information bottleneck theory [32] provides a normative framework for formalizing the notion of a relational bottleneck. Consider an information processing system that receives an input signal  $X$  and aims to predict a target signal  $Y$ .  $X$  is processed to generate a compressed representation  $Z = f(X)$  (the ‘bottleneck’), which is then used to predict  $Y$ . At the heart of information bottleneck theory is the idea of ‘minimal-sufficiency’.  $Z$  is *sufficient* for predicting  $Y$  if it contains all the information  $X$  encodes about  $Y$ . That is,  $I(Z; Y) = I(X; Y)$ , where  $I(\cdot; \cdot)$  is the mutual information. If  $Z$  is sufficient, then we write  $X \rightarrow Z \rightarrow Y$ , meaning that  $Y$  is conditionally independent of  $X$  given the compressed representation  $Z$ .  $Z$  is *minimal-sufficient* if it is sufficient for  $Y$  and does not contain any extraneous information about  $X$  which is not relevant to predicting  $Y$ . That is,  $I(X; Z) \leq I(X; \tilde{Z})$  for any other sufficient compressed representation  $\tilde{Z}$ .

Achieving maximum compression while retaining as much relevant information as possible is a trade-off. It is captured by the information bottleneck objective,

$$\text{minimize } \mathcal{L}(Z) = I(X; Z) - \beta I(Z; Y). \quad (1)$$

This objective reflects the tension between compression – which favors discarding information as captured by the first term – and the preservation of relevant information in  $Z$ , captured by the second term. The parameter  $\beta$  controls this trade-off.

While this objective is well-defined when the joint distribution  $(X, Y)$  is known, obtaining a minimal-sufficient compressed representation from data is, in general, very challenging for the high-dimensional signals that are often of interest. However, it may be possible to implicitly enforce a desirable information bottleneck for a large class of tasks through architectural inductive biases.

In particular, we hypothesize that human cognition has been fundamentally optimized for tasks that are relational in nature. We define a ‘relational task’ as any task for which there exists a minimal-

sufficient representation  $R$  that is *purely relational*. Suppose the input signal represents a set of objects,

$$X = (x_1, \dots, x_N). \quad (2)$$

A relational signal is a signal of the form,

$$R = \{r(x_i, x_j)\}_{i \neq j} = \{r(x_1, x_2), r(x_1, x_3), \dots, r(x_{N-1}, x_N)\}, \quad (3)$$

where  $r(x_i, x_j)$  is a learned relation function that satisfies certain key relational properties (e.g., transitivity). One type of operation that satisfies the relevant properties is inner products of the form  $\langle \phi(x_i), \psi(x_j) \rangle$ . Let  $\mathcal{R}$  be the class of all possible relational representations of the input signal  $X$ . In a relational task, there exists  $R \in \mathcal{R}$  which is sufficient for predicting  $Y$  (i.e.,  $X \rightarrow R \rightarrow Y$ ).

A relational bottleneck is any mechanism that restricts the space of all possible compressed representations to be a subset of the relational signals  $\mathcal{R}$ . This gives the model a smaller space of possible compressed representations over which it must search. This space of compressed representations  $\mathcal{R}$  is guaranteed to contain a minimal-sufficient representation for the task and excludes many representations that encode extraneous information about  $X$ , promoting efficient learning of relational abstractions.

## The relational bottleneck in neural architectures

Figure 2 (Key Figure) depicts three neural architectures that implement the relational bottleneck through the use of architectural inductive biases. Here, we discuss how the distinct mechanisms employed by these models implement the same underlying principle. In particular, a common aspect of all three architectures is the use of inner products to represent relations, which ensures that the resulting representations are genuinely relational. In each case, we also contrast these architectures with closely related approaches that do *not* incorporate a relational bottleneck, emphasizing how this key architectural feature gives rise to the data-efficient induction of abstractions.

### Emergent symbol binding

We first consider the Emergent Symbol Binding Network (ESBN) (Figure 2a) [33]. The ESBN is a deep neural network architecture, augmented by an external memory, that was inspired by the notion of role-filler variable binding in cognitive models of relational reasoning [36, 37, 38]. In those models, relational reasoning is supported by the existence of separately represented ‘roles’, which capture information about abstract variables, and ‘fillers’, which capture information about concrete entities to which those variables are bound. Previous work has focused on how these roles and fillers can be dynamically bound in neural circuits. However, the role and filler representations themselves were typically pre-specified by the modeler, leaving open the question of how these representations might emerge from experience.

The ESBN adopts this key idea of separate role and filler representations, but integrates them into a larger system that can be trained end-to-end, averting the need to pre-specify those representations. The ESBN contains three major components: 1) a feedforward encoding pathway (‘Encoder’ in Figure 2a), which generates object embeddings from high-dimensional perceptual inputs, 2) a recurrent controller (‘Controller’ in Figure 2a), which operates over learned representations of abstract task-relevant variables, without direct access to the object embeddings, and 3) an external memory system responsible for binding and associating representations between these two pathways. The ESBN processes perceptual observations sequentially. For each observation, a pair of embeddings is added to memory, one from the perceptual pathway (referred to as a *key*), and one from the control pathway (referred to as a *value*)<sup>1</sup>. To read from this memory, the object embedding for the current observation (referred to as a *query*) is compared to all of the keys in memory via an inner product, yielding a set of scores (one for each key) that govern the retrieval of the associated values in the abstract control pathway.

<sup>1</sup>Note that the use of the terms ‘key’ and ‘value’ here is reversed relative to the original paper [33] in order to be more consistent with their usage in describing the CoRelNet and Abstractor architectures.

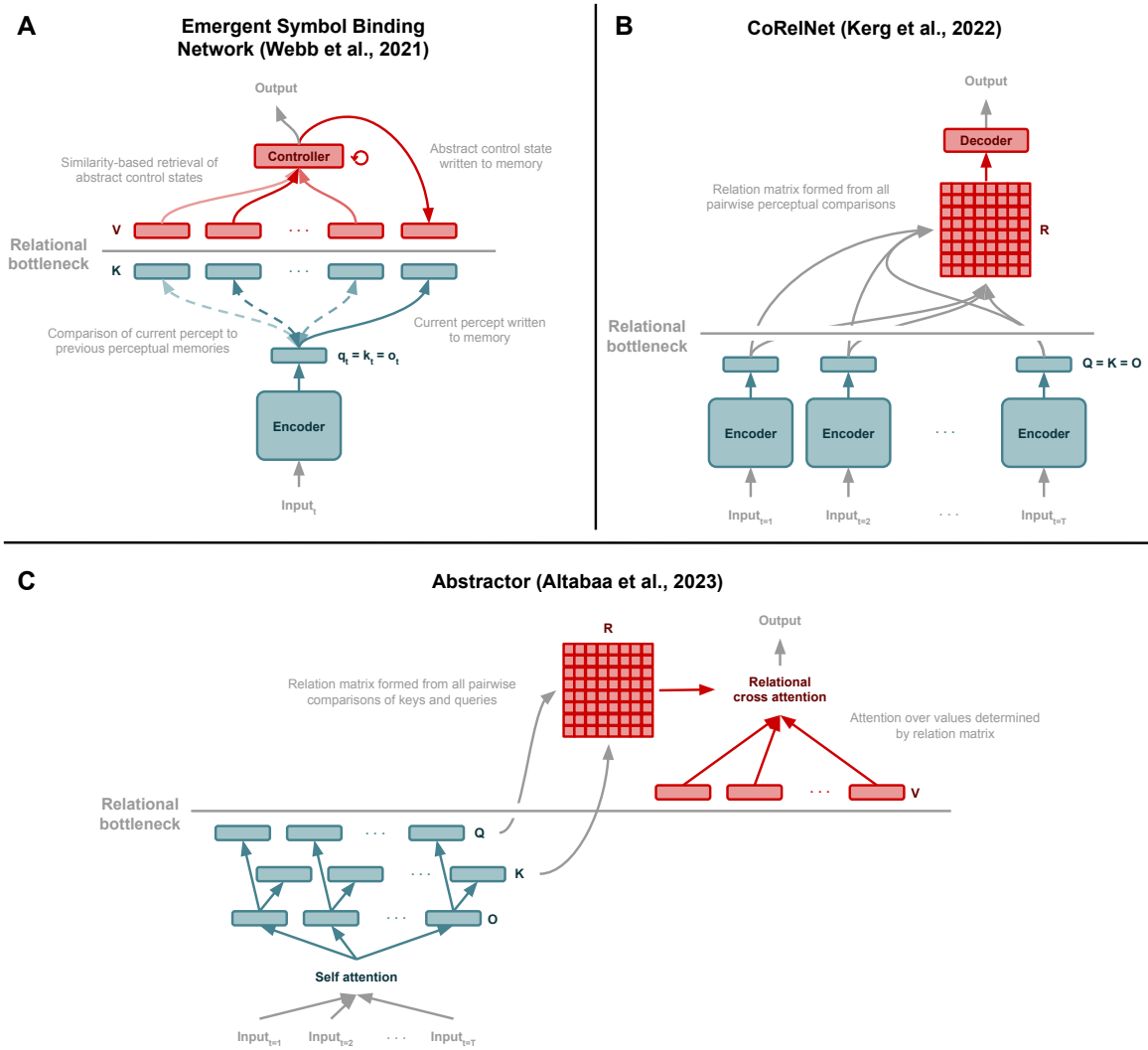


Figure 2: **Implementing the relational bottleneck.** Three neural architectures that implement the relational bottleneck. (a) Emergent Symbol Binding Network (ESBN) [33]. (b) Compositional Relation Network (CoRelNet) [34]. (c) Abstractor [35]. In all cases, high-dimensional inputs (e.g., images) are processed by a neural encoder (e.g., a convolutional network), yielding a set of object embeddings  $O$ . These are projected to a set of keys  $K$  and queries  $Q$ , which are then compared yielding a relation matrix  $R$ , in which each entry is an inner product between a query and key. Abstract values  $V$  are isolated from perceptual inputs (the core feature of the relational bottleneck), and depend only on the relations between them.

Importantly, in this retrieval operation, the control pathway does not gain access to the *content* of the representations in the perceptual pathway. Instead, the interaction is mediated only by the *comparison* of perceptual representations with each other. The ESBN thus implements the relational bottleneck as an architectural prior, separating the learning and use of abstract representations by the controller from the embeddings of perceptual information. Thanks to this design feature, the ESBN is capable of rapidly learning relational patterns (such as the identity rules displayed in Figure 1), and generalizing them to **out-of-distribution** inputs (e.g., to previously unseen shapes) [33]. Critically, it can be shown to use precisely the same representation for a given role, irrespective of filler, thus exhibiting a critical feature of abstract, symbolic processing [24]. In this sense, the representations in the model’s control pathway can be viewed as a form of learned ‘symbols’.

It is instructive to compare this model with similar approaches that do not implement the relational bottleneck. The ESBN is part of a broader family of neural network architectures that use content-addressable **external memory** – a separate store of information with which a neural network can interact via learnable read and write operations [39, 40, 41]. Notably, these read and write operations typically rely on a similarity computation (based on inner products). These have often been cast as simplified implementations of the brain’s episodic memory system [42, 43]. Standard external memory architectures do not typically isolate the control and perceptual pathways. Instead, perceptual inputs are passed directly to a central controller, which is then responsible for writing to and reading from a single, monolithic memory. Though it is possible for a role-filler structure to emerge in these systems given a sufficiently large amount of training data [44], they take much longer to learn relational tasks (requiring approximately an order of magnitude more training data), and do not display the same degree of generalization [33]. Thus, although external memory plays an important role in the ESBN architecture, the presence of external memory alone is insufficient to implement a relational bottleneck. Rather, it is the *isolation* of the perceptual and abstract processing components from one another that implements the relational bottleneck. Furthermore, as we illustrate in the following sections, it is possible to achieve this isolation without the use of external memory.

## Relation matrices

An alternative approach to implementing the relational bottleneck is illustrated by the Compositional Relation Network (CoRelNet) (Figure 2b) [34]. In that approach, a set of perceptual observations are first processed by an encoder, yielding a sequence of object embeddings. A relation matrix is then computed over all pairs of objects, in which each entry consists of the inner product between a pair of object embeddings. Finally, this relation matrix is passed to a downstream decoder network (the architecture of this network can vary, e.g., using a multilayer perceptron or transformer). This decoder is subject to a relational bottleneck, in that it only has access to the relation matrix, and does not have direct access to the object embeddings. As with the ESBN, this relational bottleneck enables CoRelNet to rapidly learn and systematically generalize relational patterns.

CoRelNet can be viewed as a feedforward, parallelized implementation of the sequential process (of encoding and similarity-based retrieval from external memory) carried out by the ESBN. This results in performance benefits, as CoRelNet does not suffer from the vanishing gradient problem that is a challenge for recurrent neural networks used to implement such sequential processing [45]. It also makes the key relational inductive bias underlying the ESBN more explicit. The ESBN’s memory retrieval procedure, in which the current observation is compared to the entries in memory, can be interpreted as computing a single row of the relation matrix. In both architectures, downstream processing is constrained so as to depend only on this relation matrix, though the details of this dependency differ.

Here too, a useful contrast can be made with related architectures that do not incorporate a relational bottleneck. In particular, architectures such as the Relation Net [46] (see [47] for related approaches) explicitly perform a comparison between each pair of inputs, leading to improved performance in relational tasks. However, whereas CoRelNet represents pairwise relations using inner products, the Relation Net utilizes generic neural network components (e.g., multilayer perceptrons) that are learned in a task-dependent manner. While this is in principle more flexible, it does not constrain the network to learn representations that *only* capture relational information. As a consequence, this architecture is susceptible to learning shortcuts consistent with the training data (i.e., overfitting to perceptual details), compromising its ability to reliably generalize learned relations to out-of-distribution inputs [48, 33, 49]. This is in contrast to the inner product operation employed by the ESBN and CoRelNet, which is inherently relational, and therefore guarantees that downstream processing is based only on relations.

## Relational attention

The recently proposed Abstractor architecture (Figure 2c) [35] illustrates how the relational bottleneck can be implemented within the broader framework of attention-based architectures (including the Transformer [50]). The Abstractor is built on a novel attention operation termed *relational cross-attention*. In this operation, a set of object embeddings (which may be produced by an encoder given perceptual observations) is converted to form keys and queries, using separate linear projections. A relation matrix is then computed, in which each entry corresponds to the inner product between a

query and key. The relation matrix is used to attend over a set of learned values, which reference objects but are independent of their attributes.

Relational cross-attention can be contrasted with the standard forms of attention employed in Transformers: self-attention and cross-attention. In self-attention, the same set of object embeddings are used to generate keys, queries, and values. In cross-attention, object embeddings are used to generate keys and values, and queries are generated by a separate decoder network. In both cases, the values over which attention is performed are based directly on the object embeddings, and the information contained in these embeddings is therefore passed on for downstream processing (thus contravening the relational bottleneck). By contrast, in *relational* cross-attention, keys and queries are generated from object embeddings, but a separate set of learned vectors are used as values. As in the ESN, these values can be viewed as learned ‘symbols’, in the sense that they are isolated from the perceptual content of the objects with which they are associated.

This implementation of the relational bottleneck yields the same benefits observed in others: the Abstractor learns relational patterns faster than the Transformer, and displays better out-of-distribution generalization of those patterns [35]<sup>2</sup>. The Abstractor also has a few advantages relative to existing implementations of the ESN and CoRelNet. Because the relation matrix is computed using separate key and query projections, the Abstractor is capable of representing asymmetric relations (e.g., can capture the difference in meanings between ‘A is greater than B’ and ‘B is greater than A’). In addition, multi-headed relational cross-attention enables the Abstractor to model multi-dimensional relations. As proposed, ESN and CoRelNet are limited to relations along a single feature dimension only. Finally, similar to Transformers, the Abstractor is a *generative* architecture, whereas the ESN and CoRelNet are purely discriminative<sup>3</sup>. This enables the Abstractor to perform a broader range of tasks, including the sequence-to-sequence tasks that are common in natural language processing.

As the examples we have considered illustrate, the relational bottleneck can be implemented in a diverse range of architectures, each with their own strengths and weaknesses. In each case, the inclusion of a relational bottleneck enables rapid learning of relations without the need for pre-specified relational primitives. In the remainder of the review, we discuss the implications of this approach for models of cognition, and consider how the relational bottleneck may relate to the architecture of the human brain.

---

## Box 2: Neuro-symbolic modeling approaches

Many approaches have been proposed for hybrid systems that combine aspects of both neural and symbolic computing. Early work in this area focused on incorporating a capacity for variable-binding – a key property of symbolic systems – into connectionist systems. Notable examples of this approach include binding-by-synchrony [37], tensor product variable-binding [36], and BoltzCONS [52]. A number of vector symbolic architectures have since been proposed that build on the tensor product operation, but enable more elaborate symbolic structures to be embedded in a vector space of fixed dimensionality [53, 54, 55, 56]. These approaches have all generally relied on the use of pre-specified symbolic primitives.

More recently, hybrid systems have been developed that combine deep learning with symbolic programs. In this approach, deep learning components are typically employed to translate raw perceptual inputs, such as images or natural language, into symbolic representations, which can then be processed by traditional symbolic algorithms [57, 58, 59, 60, 61]. This approach is complemented by recent neuro-symbolic approaches to probabilistic program induction, in which symbolic primitives are pre-specified (following earlier symbolic-connectionist modeling efforts), and then deep learning is used to assemble these primitives into programs [28].

---

<sup>2</sup>Although, as noted in the introduction, there is evidence that the standard Transformer architecture can learn to perform relational tasks (e.g., in the case of large language models [18]), this requires considerable amounts of data. Experiments comparing the standard Transformer architecture with the various implementations of the relational bottleneck highlighted above suggest that the latter may be substantially more data efficient, though this remains to be demonstrated at scale, and for the full range of tasks over which Transformers have been applied.

<sup>3</sup>It should be noted that these are not fundamental limitations of the ESN and CoRelNet architectures. For instance, both of these architectures can be modified so as to employ separate key and query embeddings, enabling asymmetric relations to be modeled [34]. Furthermore, an alternative implementation of the ESN has been proposed that can perform generative tasks [51].

An alternative approach (which might also be viewed as neuro-symbolic in some sense) involves the integration of key features of symbolic computing within the framework of end-to-end trainable neural systems. Examples of this approach include neural production systems [62], graph neural networks [47], discrete-valued neural networks [63], and efforts to incorporate tensor product representations into end-to-end systems [64, 65]. The relational bottleneck falls into this broad category, as it incorporates key elements of symbolic computing – variable-binding and relational representations – into fully differentiable neural systems that can be trained end-to-end without the need for pre-specified symbolic primitives. Relative to these other approaches, the primary innovation of the relational bottleneck framework is the emphasis on architectural components that promote the development of genuinely relational representations.

---

## The relational bottleneck in the mind and brain

### Modeling the development of counting: a case study in learning abstractions

A core requirement for cognitive models of abstract concept acquisition is to account for the timecourse of acquisition during human development. A useful case study can be found in the early childhood process of learning to count [66, 67, 68]. Children typically learn to recite the count sequence (i.e. ‘one, two, three,...’ etc.) relatively early, but their ability to use this knowledge to count objects then proceeds in distinct stages. Each stage is characterized by the ability to reliably count sets up to a certain size (i.e., first acquiring the ability to reliably count only single objects, then to count two objects, and so on). Around the time that children learn to count sets of five, an inductive transition occurs, in which children rapidly learn to counts sets of increasing size. It has been proposed that this transition corresponds to the acquisition of the ‘cardinality principle’ – the understanding that the last word used when counting corresponds to the number of items in a set [68].

A recent study investigated the development of counting in deep neural network architectures [69]. These included the ESN, the Transformer, and long short-term memory (LSTM) [70] (a type of recurrent neural network). Each architecture displayed a distinct developmental timecourse. The Transformer displayed a roughly linear timecourse, taking approximately the same amount of time to master each number. The LSTM displayed an exponentially increasing timecourse, taking more time to learn each new number. Only the ESN displayed a human-like inductive transition, gradually learning to count each number from one to four, and then rapidly acquiring the ability to count higher after learning to count to five. This was due to the ability of the ESN to learn a procedure over the representations in its control pathway that was abstracted away from the specific numbers in the count sequence (represented in the model’s perceptual pathway), allowing it to rapidly and systematically generalize between numbers. This case study illustrates how the relational bottleneck can emulate a human-like developmental trajectory for learning abstract concepts.

### Cognitive models of analogical reasoning

The relational bottleneck also has some important parallels with cognitive models of analogical reasoning. In particular, both approaches afford an especially important role to *patterns of similarity*. In traditional symbolic models, this typically takes the form of literal identity between symbols [71]. However, more recent models employ a graded measure of similarity that can be easily applied to distributed representations, such as those derived from deep learning systems (e.g., word or image embeddings) [72, 73, 74]. In those models, a similarity matrix is computed, which is then used to identify a mapping from the elements in one situation to the elements in another situation. This use of similarity matrices has a close connection to the relation matrices (both of which are based on inner products) employed explicitly in architectures such as CoRelNet and the Abstractor, and implicitly in the retrieval operation of the ESN. This suggests the intriguing possibility that these architectures, aided by the inductive bias of the relational bottleneck, may learn to implement a procedure similar to the mapping algorithm proposed by cognitive models of analogy.



## Capacity limits and the curse of compositionality

The relational bottleneck principle may also help to explain the limited capacity of some cognitive processes (e.g., working memory) [75]. Recent work has demonstrated that human-like capacity limits naturally emerge in an architecture that implements the relational bottleneck [76]. In that architecture, two separate representational pools (each representing distinct feature spaces, e.g., color and location) interact via a dynamic variable-binding mechanism (in that case, implemented using rapid Hebbian learning). This architecture is conceptually similar to the ESN, but is subject to classical efficient coding constraints—that is, limits not only on the amount of available data, but also time and memory available for optimizing a loss function. This mechanism, which is intimately related to classic neural network models of rapid learning and memory retrieval [77], enables the model to flexibly construct compositional representations (e.g., representing a visual scene by binding together spatial locations and visual features). However, this flexibility comes at the cost of relying on compositional representations that, by definition, are shared across many different, potentially competing processes (an instance of the general relationship between shared representations and cognitive capacity [78]). The model quantitatively captures capacity limits observed in three distinct cognitive domains: working memory [75], subitizing (the ability to rapidly identify the number of items in a display) [79], and absolute judgment (the ability to correctly label specific feature values such as pitch or loudness) [80].

## Brain mechanisms supporting the relational bottleneck

We close by considering how the relational bottleneck might relate to the architecture of the human brain. A central element of this framework is the presence of segregated systems for representing abstract vs. perceptual information (i.e., abstract values vs. perceptual keys/queries in the ESN or Abstractor). A large body of findings from cognitive neuroscience suggests the presence of distinct neocortical systems for representing abstract structure (e.g., of space or events) vs. concrete entities (e.g., people or places), located in the parietal and temporal cortices respectively [81, 82, 83, 84, 85]. This factorization has also been explored in a number of recent computational models [86, 87, 88].

However, this segregation raises the question of how representations in these distinct neocortical systems are flexibly bound together. Though many proposals have been made for how the brain might solve this variable-binding problems (see Box 2), one intriguing possibility involves the episodic memory system [42]. A common view holds that episodic memory is supported by rapid synaptic plasticity in the hippocampus, which complements slower statistical learning in the neocortex [43, 89]. According to this view, episodes are encoded in the hippocampus by the rapid *binding* of features that co-occur within an episode, while the features themselves are represented in neocortical systems. This same mechanism could in principle support an architecture similar to the ESN, by enabling rapid binding of abstract and perceptual neocortical representations. This is in fact very similar to models of cognitive map learning, according to which distinct representations of structural vs. sensory information, corresponding to the medial vs. lateral entorhinal cortices (often viewed as extensions of the parietal and temporal neocortical systems referenced above), are bound together by rapidly formed conjunctive representations in the hippocampus [90].

That said, the extent to which variable-binding relies on the hippocampus remains an open question. Some lesion evidence suggests that hippocampal damage does not lead to impairments of abstract reasoning [91] (see Box 3 for further discussion). Other alternatives are that variable-binding may be supported by other structures capable of rapid synaptic plasticity (e.g., the cerebellum, which has been increasingly implicated in higher cognitive functions [92, 93, 94]), or by other structures (such as the prefrontal cortex) that use other mechanisms for binding (such as selective attention [95] or working memory gating [96]). The latter possibilities are consistent with findings that prefrontal damage often leads to severe deficits in abstract reasoning tasks [97, 98], and prefrontal activity is frequently implicated in neuroimaging studies of abstract reasoning [99, 100]. However, this may also reflect the role of prefrontal cortex in *representing* abstract structure (along with the parietal system described above), rather than the *binding* of that structural information to concrete content. Of course, it is also possible that variable-binding is supported by a collection of distinct mechanisms, rather than a single mechanism alone. These are all important questions for future work that we hope will be usefully guided by the formalisms and computational models reviewed here.

---

### Box 3: Episodic memory and the relational bottleneck

The proposal that episodic memory (EM) plays a crucial role in abstract reasoning may seem to be at odds with conventional wisdom for several reasons. First, the capacity for abstraction may be assumed to fall more naturally within the province of semantic memory, which is generally assumed to encode the abstract (e.g., statistical) structure of relationships among concepts [43, 89]. The proposal considered here is not that EM *represents* such structure, but rather that it is used to apply structural information (e.g., roles) to specific instances (e.g., fillers) by serving as a binding mechanism.

Another concern might be that reasoning processes are generally associated with working memory (WM) function [101, 95] rather than EM. However, a growing body of recent findings have suggested the potential involvement of EM in tasks that are traditionally associated with WM [102, 103, 104]. Furthermore, the functional properties of EM are well suited to perform the variable-binding operation that plays a critical role in the relational bottleneck framework. In traditional accounts of EM, rapid hippocampal plasticity serves to bind together the features of an episode, but this same mechanism is in principle capable of binding together abstract and perceptual representations (such as the key and value representations in the ESNB) in the service of an ongoing reasoning process.

As noted in the text, there is some evidence that could be taken as evidence against this account: lesion studies suggesting that hippocampal damage, which leads to severe EM deficits, does not lead to comparably severe deficits in abstract reasoning [91], whereas reasoning deficits often arise from damage to prefrontal cortex [97, 98]. It is of course possible that both hippocampal and prefrontal mechanisms contribute to variable-binding in the healthy brain, but that prefrontal mechanisms alone can support variable-binding in the event of hippocampal damage. However, an alternative possibility is that EM-like processes – i.e., the rapid encoding of arbitrary but durable associations subject to similarity-based retrieval – may be subserved by other brain regions not traditionally associated with EM, such as the prefrontal cortex, cerebellum, or other structures. From this perspective, the relational bottleneck framework points to a number of intriguing directions for future research concerning the nature of EM and its relationship to the capacity for abstraction.

---

## Concluding remarks and future directions

The human mind has a remarkable ability to acquire abstract relational concepts from relatively limited and concrete experience. Here, we have proposed the relational bottleneck as a functional principle that may explain how the human brain accomplishes such data-efficient abstraction, and highlighted recently proposed computational models that implement this principle. We have also considered how the principle relates to a range of cognitive phenomena, and how it might be implemented by the mechanisms of the human brain.

It should be noted that the framework reviewed here is not necessarily at odds with the existence of certain forms of domain-specific innate knowledge. In particular, a range of evidence from developmental psychology has suggested that humans possess certain ‘core knowledge’ systems, such as an innate capacity to represent objects [105, 106, 107]. These findings have motivated the development of neuro-symbolic models endowed with these innate capacities [108], although it is also possible that these findings may ultimately be accounted for by the inclusion of additional inductive biases into connectionist systems, such as mechanisms for object-centric visual processing [109, 110, 111, 112] (which have also been combined with the relational bottleneck [113]). Critically, however, it is important to emphasize that the relational bottleneck is, in principle, orthogonal to questions about these domain-specific capacities, and is focused instead on explaining the induction of abstract, domain-general concepts and relations.

There are a number of important avenues for further developing the relational bottleneck framework (see Outstanding Questions). Further work is needed to integrate the relational bottleneck with a broader range of cognitive processes relevant to abstraction, including attentional processes [114] and semantic cognition [115]. Additionally, much work has suggested that human reasoning is not purely relational, but instead depends on a mixture of concrete and abstract influences [116, 117, 118, 119]. This suggests the potential value of a more graded formulation that controls the amount

of non-relational information allowed to pass through the bottleneck. Finally, the human capacity for abstraction surely depends not only on architectural biases such as those that we have discussed here, but also on the rich educational and cultural fabric that allows us to build on the abstractions developed by others [120]. In future work, it will be important to explore the interaction between education, culture and relational inductive biases.

## Outstanding Questions

- Human reasoners often display so-called ‘content effects’, in which abstract reasoning processes are influenced by the specific content under consideration (and therefore are not purely abstract or relational). Can a more graded version of the relational bottleneck capture these effects, while preserving a capacity for relational abstraction?
- How can other cognitive processes (perception, attention, memory, etc.) be integrated with the relational bottleneck?
- How is the relational bottleneck implemented in the brain? To what extent does this rely on mechanisms responsible for episodic memory, attentional mechanisms, and/or other mechanisms that remain to be identified? What role do the hippocampus, prefrontal cortex, and/or other structures play in these computations?
- How do architectural biases toward relational processing interact with cultural sources of abstraction (e.g., formal education)?

## Glossary

### Connectionism

A modeling framework in cognitive science that emphasizes the emergence of complex cognitive phenomena from the interaction of simple, neuron-like elements organized into networks, in which connections are formed through learning.

### Empiricism

An epistemological view according to which knowledge is ultimately derived from experience. Often contrasted with nativism.

### Episodic memory

A form of memory in which arbitrary, but durable, associations can be rapidly formed. Often thought to be implemented by hippocampal mechanisms for rapid synaptic plasticity and similarity-based retrieval.

### External memory

In the context of neural networks, an approach that combines these with separate external stores of information, typically with learnable mechanisms for writing to and reading from these stores, and in which retrieval is usually similarity-based (i.e., ‘content-addressable’). Often used to implement a form of episodic memory.

### Inductive bias

An assumption made by a machine learning model about the distribution of the data. In deep learning models, this often takes the form of architectural features that bias learning toward certain (typically desirable) outcomes. Genetically pre-configured aspects of brain structure can be viewed as a form of inductive bias.

## Out-of-distribution generalization

In machine learning, generalization to a distribution that differs from the distribution observed during training.

## Nativism

The view that certain concepts and mental capacities are innate rather than learned from experience. Often contrasted with empiricism.

## Declaration of interests

The authors declare no competing interests.

## References

- [1] Descartes, R. (1988). “Rules for the Direction of our Native Intelligence”. Descartes: Selected Philosophical Writings, J. Cottingham, R. Stoothoff, and D. Murdoch, ed. (Cambridge University Press).
- [2] Locke, J. (1894). An essay concerning human understanding. A.C. Fraser, ed. (Oxford: Clarendon Press).
- [3] Leibniz, G. (1996). New Essays on Human Understanding. P. Remnant and J. Bennett, ed. (New York: Cambridge University Press).
- [4] Chomsky, N. (1980). A review of BF Skinner’s Verbal Behavior. The Language and Thought Series, 48–64.
- [5] Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7, 155–170.
- [6] Holyoak, K.J. (2012). “Analogy and Relational Reasoning”. The Oxford Handbook of Thinking and Reasoning, K.J. Holyoak and R.G. Morrison, ed. (Oxford University Press), pp. 234–259.
- [7] Cattell, R.B. (1971). Abilities: Their structure, growth, and action (Houghton Mifflin).
- [8] Snow, R.E., Kyllonen, P.C., Marshalek, B., et al. (1984). The topography of ability and learning correlations. *Advances in the Psychology of Human Intelligence* 2, 103.
- [9] Penn, D.C., Holyoak, K.J., and Povinelli, D.J. (2008). Darwin’s mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences* 31, 109–130.
- [10] McClelland, J.L. and Rumelhart, D.E. (1989). Explorations in parallel distributed processing: A handbook of models, programs, and exercises (MIT Press).
- [11] Elman, J.L. (1990). Finding structure in time. *Cognitive Science* 14, 179–211.
- [12] McClelland, J.L. and Rogers, T.T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience* 4, 310–322.
- [13] McClelland, J.L. et al. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14, 348–356.
- [14] Brown, T. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877–1901.
- [15] Wei, J. et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Survey Certification. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=yzkSU5zdwD>.
- [16] Piantadosi, S. (2023). Modern language models refute Chomsky’s approach to language. *Lingbuzz Preprint*, lingbuzz 7180.
- [17] Bubeck, S. et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- [18] Webb, T., Holyoak, K.J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*. URL: <https://doi.org/10.1038/s41562-023-01659-w>.
- [19] Griffiths, T.L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences* 24, 873–883.
- [20] Frank, M.C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*. URL: <https://doi.org/10.1016/j.tics.2023.08.007>.

- [21] Newell, A., Simon, H.A., et al. (1972). Human problem solving vol. 104. (Prentice-hall Englewood Cliffs, NJ).
- [22] Fodor, J.A. (1975). The language of thought vol. 5. (Harvard University Press).
- [23] Anderson, J.R. (1996). ACT: A simple theory of complex cognition. *American Psychologist* 51, 355.
- [24] Fodor, J.A. and Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 3–71.
- [25] Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. *Science* 350, 1332–1338.
- [26] Lake, B.M. et al. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences* 40, e253.
- [27] Rule, J.S., Tenenbaum, J.B., and Piantadosi, S.T. (2020). The child as hacker. *Trends in Cognitive Sciences* 24, 900–915.
- [28] Ellis, K. et al. (2021). Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. *Proceedings of the 42nd ACM Sigplan International Conference on Programming Language Design and Implementation*. URL: <https://doi.org/10.1145/3410302>.
- [29] Dehaene, S. et al. (2022). Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*. URL: <https://doi.org/10.1016/j.tics.2022.06.010>.
- [30] Yang, Y. and Piantadosi, S.T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences* 119, e2021865119.
- [31] Quilty-Dunn, J., Porot, N., and Mandelbaum, E. (2022). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*. URL: <https://doi.org/10.1017/S0140525X22002849>.
- [32] Tishby, N., Pereira, F.C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.
- [33] Webb, T.W., Sinha, I., and Cohen, J.D. (2021). Emergent Symbols through Binding in External Memory. *9th International Conference on Learning Representations (ICLR)*. URL: <https://openreview.net/forum?id=LSFCEb3GYU7>.
- [34] Kerg, G. et al. (2022). On neural architecture inductive biases for relational tasks. *arXiv preprint arXiv:2206.05056*.
- [35] Altabaa, A. et al. (2023). Abstractors: Transformer Modules for Symbolic Message Passing and Relational Reasoning. *arXiv preprint arXiv:2304.00195*.
- [36] Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46, 159–216.
- [37] Hummel, J.E. and Holyoak, K.J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104, 427.
- [38] Marcus, G.F. (2001). *The algebraic mind: Integrating connectionism and cognitive science* (MIT Press).
- [39] Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.
- [40] Graves, A. et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature* 538, 471–476.
- [41] Pritzel, A. et al. (2017). Neural episodic control. *34th International Conference on Machine Learning*. URL: <https://proceedings.mlr.press/v70/pritzel17a.html>.
- [42] Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology* 53, 1–25.
- [43] McClelland, J.L., McNaughton, B.L., and O’Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102, 419.
- [44] Chen, C. et al. (2021). Learning to perform role-filler binding with schematic knowledge. *PeerJ* 9, e11046.
- [45] Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 107–116.
- [46] Santoro, A. et al. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems* 30, 4974–4983.

- [47] Battaglia, P.W. et al. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.
- [48] Kim, J., Ricci, M., and Serre, T. (2018). Not-So-CLEVR: learning same-different relations strains feedforward neural networks. *Interface Focus* 8, 20180011.
- [49] Ichien, N. et al. (2021). Visual analogy: Deep learning versus compositional models. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. URL: <https://escholarship.org/uc/item/36k485sw>.
- [50] Vaswani, A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30, 6000–6010.
- [51] Sinha, I., Webb, T.W., and Cohen, J.D. (2020). A memory-augmented neural network model of abstract rule learning. arXiv preprint arXiv:2012.07172.
- [52] Touretzky, D.S. (1990). BoltzCONS: Dynamic symbol structures in a connectionist network. *Artificial Intelligence* 46, 5–46.
- [53] Plate, T.A. (1995). Holographic reduced representations. *IEEE Transactions on Neural networks* 6, 623–641.
- [54] Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation* 1, 139–159.
- [55] Eliasmith, C. et al. (2012). A large-scale model of the functioning brain. *Science* 338, 1202–1205.
- [56] Schlegel, K., Neubert, P., and Protzel, P. (2022). A comparison of vector symbolic architectures. *Artificial Intelligence Review* 55, 4523–4555.
- [57] Andreas, J. et al. (2016). Learning to compose neural networks for question answering. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. URL: <https://aclanthology.org/N16-1181>.
- [58] Johnson, J. et al. (2017). Inferring and executing programs for visual reasoning. *Proceedings of the IEEE International Conference on Computer Vision*, 2989–2998.
- [59] Yi, K. et al. (2018). Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. *Advances in Neural Information Processing Systems* 31, 1039–1050.
- [60] Mao, J. et al. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *7th International Conference on Learning Representations (ICLR)*. URL: <https://openreview.net/forum?id=rJgMlhRctm>.
- [61] Nye, M. et al. (2020). Learning compositional rules via neural program synthesis. *Advances in Neural Information Processing Systems* 33, 10832–10842.
- [62] Goyal, A. et al. (2021). Neural production systems. *Advances in Neural Information Processing Systems* 34, 25673–25687.
- [63] Liu, D. et al. (2021). Discrete-valued neural communication. *Advances in Neural Information Processing Systems* 34, 2109–2121.
- [64] Palangi, H. et al. (2018). Question-answering with grammatically-interpretable representations. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 5350–5357.
- [65] Jiang, Y. et al. (2021). Enriching transformers with structured tensor-product representations for abstractive summarization. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. URL: <https://aclanthology.org/2021.naacl-main.381>.
- [66] Wynn, K. (1992). Children’s acquisition of the number words and the counting system. *Cognitive Psychology* 24, 220–251.
- [67] Carey, S. (2001). Cognitive foundations of arithmetic: Evolution and ontogenesis. *Mind & Language* 16, 37–55.
- [68] Sarnecka, B.W. and Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition* 108, 662–674.
- [69] Dulberg, Z., Webb, T., and Cohen, J. (2021). Modelling the development of counting with memory-augmented neural networks. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. URL: <https://escholarship.org/uc/item/34z9j7q0>.
- [70] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* 9, 1735–1780.
- [71] Falkenhainer, B., Forbus, K.D., and Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41, 1–63.

- [72] Lu, H., Wu, Y.N., and Holyoak, K.J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences* 116, 4176–4181.
- [73] Lu, H., Ichien, N., and Holyoak, K.J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review*.
- [74] Webb, T.W. et al. (2023). Zero-shot visual reasoning through probabilistic analogical mapping. *Nature Communications* 14, 5144.
- [75] Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 81.
- [76] Frankland, S.M., Webb, T., and Cohen, J.D. (2021). No coincidence, George: Capacity-limits as the Curse of Compositionality. *PsyArXiv preprint*. URL: [psyarxiv.com/cjuxb](https://psyarxiv.com/cjuxb).
- [77] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* 79, 2554–2558.
- [78] Musslick, S. and Cohen, J.D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences* 25, 757–775.
- [79] Mandler, G. and Shebo, B.J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General* 111, 1.
- [80] Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America* 24, 745–749.
- [81] Mishkin, M., Ungerleider, L.G., and Macko, K.A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences* 6, 414–417.
- [82] Goodale, M.A. and Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences* 15, 20–25.
- [83] Frankland, S.M. and Greene, J.D. (2020). Concepts and compositionality: in search of the brain’s language of thought. *Annual Review of Psychology* 71, 273–303.
- [84] Summerfield, C., Luyckx, F., and Sheahan, H. (2020). Structure learning and the posterior parietal cortex. *Progress in Neurobiology* 184, 101717.
- [85] O’Reilly, R.C., Ranganath, C., and Russin, J.L. (2022). The structure of systematicity in the brain. *Current Directions in Psychological Science* 31, 124–130.
- [86] Russin, J. et al. (2019). Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*.
- [87] O’Reilly, R.C. et al. (2021). Deep predictive learning in neocortex and pulvinar. *Journal of Cognitive Neuroscience* 33, 1158–1196.
- [88] Bakhtiari, S. et al. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems* 34, 25164–25178.
- [89] Sun, W. et al. (2023). Organizing memories for generalization in complementary learning systems. *Nature Neuroscience* 26, 1438–1448.
- [90] Whittington, J.C. et al. (2020). The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* 183, 1249–1263.
- [91] Dzieciol, A.M. et al. (2017). Hippocampal and diencephalic pathology in developmental amnesia. *Cortex* 86, 33–44.
- [92] Ravizza, S.M. et al. (2006). Cerebellar damage produces selective deficits in verbal working memory. *Brain* 129, 306–320.
- [93] D’Mello, A.M., Gabrieli, J.D., and Nee, D.E. (2020). Evidence for hierarchical cognitive control in the human cerebellum. *Current Biology* 30, 1881–1892.
- [94] McDougle, S.D. et al. (2022). Continuous manipulation of mental representations is compromised in cerebellar degeneration. *Brain* 145, 4246–4263.
- [95] Miller, E.K. and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24, 167–202.
- [96] Kriete, T. et al. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences* 110, 16390–16395.
- [97] Waltz, J.A. et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science* 10, 119–125.
- [98] Cipolotti, L. et al. (2023). Graph lesion-deficit mapping of fluid intelligence. *Brain* 146, 167–181.
- [99] Christoff, K. et al. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage* 14, 1136–1149.

- [100] Knowlton, B.J. et al. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences* 16, 373–381.
- [101] Baddeley, A. (1992). Working memory. *Science* 255, 556–559.
- [102] Hoskin, A.N. et al. (2019). Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience* 19, 338–354.
- [103] Beukers, A.O. et al. (2021). Is activity silent working memory simply episodic memory? *Trends in Cognitive Sciences* 25, 284–293.
- [104] Beukers, A. et al. (2022). When Working Memory May Be Just Working, Not Memory. *PsyArXiv preprint*. URL: <https://psyarxiv.com/jtw5p>.
- [105] Spelke, E.S. et al. (1992). Origins of knowledge. *Psychological Review* 99, 605.
- [106] Spelke, E.S. and Kinzler, K.D. (2007). Core knowledge. *Developmental Science* 10, 89–96.
- [107] Baillargeon, R. and Carey, S. (2012). “Core cognition and beyond: The acquisition of physical and numerical knowledge”. *Early childhood development and later outcome*, S.M. Pauen, ed. (Cambridge University Press), pp. 35–65.
- [108] Smith, K. et al. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in Neural Information Processing Systems* 32, 8985–8995.
- [109] Burgess, C.P. et al. (2019). Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*.
- [110] Locatello, F. et al. (2020). Object-centric learning with slot attention. *Advances in Neural Information Processing Systems* 33, 11525–11538.
- [111] Piloto, L.S. et al. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour* 6, 1257–1267.
- [112] Mondal, S.S., Webb, T., and Cohen, J.D. (2023). Learning to reason over visual objects. *11th International Conference on Learning Representations (ICLR)*. URL: [https://openreview.net/forum?id=uR6x8Be7o\\_M](https://openreview.net/forum?id=uR6x8Be7o_M).
- [113] Webb, T.W., Mondal, S.S., and Cohen, J.D. (2023). Systematic Visual Reasoning through Object-Centric Relational Abstraction. *arXiv preprint arXiv:2306.02500*.
- [114] Vaishnav, M. and Serre, T. (2022). GAMR: A guided attention model for (visual) reasoning. *11th International Conference on Learning Representations (ICLR)*. URL: <https://openreview.net/forum?id=iLMgk2IGNyv>.
- [115] Giallanza, T. et al. (2023). An Integrated Model of Semantics and Control. *PsyArXiv preprint*. URL: <https://psyarxiv.com/jq7ta/>.
- [116] Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology* 20, 273–281.
- [117] Johnson-Laird, P.N., Legrenzi, P., and Legrenzi, M.S. (1972). Reasoning and a sense of reality. *British Journal of Psychology* 63, 395–400.
- [118] Bassok, M., Chase, V.M., and Martin, S.A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology* 35, 99–134.
- [119] Goldberg, A.E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7, 219–224.
- [120] McClelland, J.L. (2022). Capturing advanced human cognitive abilities with deep neural networks. *Trends in Cognitive Sciences* 26, 1047–1050.