A Relational Inductive Bias for Dimensional Abstraction in Neural Networks

Declan Campbell [1] **Jonathan D. Cohen** [1, 2]

1 Princeton Neuroscience Institute
2 Princeton University Department of Psychology

Abstract

The human cognitive system exhibits remarkable flexibility and generalization capabilities, partly due to its ability to form low-dimensional, compositional representations of the environment. In contrast, standard neural network architectures often struggle with abstract reasoning tasks, overfitting, and requiring extensive data for training. This paper investigates the impact of the relational bottleneck—a mechanism that focuses processing on relations among inputs—on the learning of factorized representations conducive to compositional coding and the attendant flexibility of processing. We demonstrate that such a bottleneck not only improves generalization and learning efficiency, but also aligns network performance with human-like behavioral biases. Networks trained with the relational bottleneck developed orthogonal representations of feature dimensions latent in the dataset, reflecting the factorized structure thought to underlie human cognitive flexibility. Moreover, the relational network mimics human biases towards regularity without pre-specified symbolic primitives, suggesting that the bottleneck fosters the emergence of abstract representations that confer flexibility akin to symbols.

Keywords: Abstraction, Neural Networks, Reasoning

Introduction

The flexibility and generativity of human cognition has been the focus of intense research in cognitive science and machine learning. Fodor & Pylyshyn (1988) famously proposed that this depends on the property of compositionality supported by symbolic processing which, they argued, was not supported by neural network architectures. In contrast, a strong focus and much of the success of early neural network architectures (McClelland et al., 1986; Rumelhart & McClelland, 1986), as well as modern forms of deep learning (LeCun et al., 2015), has been on the ability to implement sophisticated forms of statistical learning. In addition to achieving powerful forms of function approximation, these architectures have been shown to achieve capabilities traditionally associated with symbolic processing (Rogers & McClelland, 2004), including the ability of transformer-based, large-language models to solve challenging analogical reasoning tasks (T. Webb et al., 2023). Nevertheless: a) these models fail on other forms of abstract reasoning tasks on which humans succeed (Mitchell et al., 2023); b) they require amounts of data and training that are far in excess of what people require to achieve comparable performance (Mnih et al., 2015); and c) at best, it is unclear whether they make use of the same kinds of low dimensional, highly abstract, compositional (and, in the limit, genuinely symbolic) forms of representation that Fodor and Pylyshyn had in mind.

Interestingly, it has recently been shown that introducing simple inductive biases to standard neural network architectures can promote the efficient learning of and flexible use of abstract representations and, in some cases, the emergence of genuinely symbolic forms of processing (T. W. Webb et al., 2020; Kerg et al., 2022; Altabaa et al., 2023). These inductive biases implement a form of relational bottleneck: a processing component that restricts the flow of information to relations among the inputs (T. W. Webb et al., 2023), upon which further processing is based. This isolation between processing of the data and processing that is restricted to only relational information inherent in the data forces the latter to learn abstract representations, and functions that use these, that can be generalized to new, never seen inputs that exhibit the same relations.

One corollary of the idea that a relational bottleneck can induce the learning of abstract representations is that this should be facilitated by representation of data in a compositional form — that is, factorized in a form that the relational structure is most apparent. The learning of such factorized representations are a requisite for compositional coding. One example of such factorization that has been an important goal of research on deep learning (Higgins et al., 2016; Kim & Mnih, 2018) is the learning of dimensional structure that may exist in the data (e.g., the color, size, and brightness of visual images). Here, we explore the possibility that the relational bottleneck can play an important role in promoting the learning of such representations. Specifically, we test the hypothesis that, insofar as factorized representations of distinct, task-relevant feature dimensions make learning relations easier, then imbuing a network with a relational bottleneck may put pressure on the network to learn such representations, improving generalization performance on relational tasks, and better aligning network performance with human behavioral biases.

To test this hypothesis, we focused on the learning of similarity relations using the simplest form of a relational bottleneck, which can be implemented in a contrastive network. Specifically, we implemented two standard multilayer perceptron (MLP) pathways with shared encoder weights (as input streams for the two items to be compared), that converged on a layer in which the distance (e.g., cosine similarity) was

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

computed between pairs of embeddings. This was used to evaluate the similarity between the embeddings in each input stream, that was taken as the response. The network was trained to produce the correct response (i.e., how similar the two inputs were) for each pair of stimuli. This implementation of the relational bottleneck is comparable to the CoRelNet architecture described by Kerg et al. (2022). We compared the performance of this network, and the representations it learned, with a standard feedforward neural network that lacked a relational bottleneck (i.e., the direct similarity computation).

We carried out two sets of simulations to evaluate these networks. In the first, we evaluated the extent to which they developed orthogonal (factorized) representations of distinct feature dimensions in the embedding layers, using a training dataset comprised of features that varied along two orthogonal dimensions. In the second set of simulations, we extended the approach to a richer feature space of geometric forms previously generated using symbolic primitives, and tested the extent to which the networks captured the difference between performance of humans and non-humans that has previously been explained using models explicitly imbued with symbolic primitives (Dehaene et al., 2022; Sablé-Meyer et al., 2021, 2022).

Relational Bottleneck and Dimensional Representations

Our initial investigation focused on a simple similarity judgment task, to evaluate the extent to which a relational bottleneck induced the learning of factorized representations of two orthogonal feature dimensions in its embeddings. In contrast to existing unsupervised methods for learning factorized representations (Higgins et al., 2016; Kim & Mnih, 2018), we used a contrastive loss function that measures the relation (distance) between pairs of inputs within the metric space defined by the dataset's latent features. We show that this approach encourages the model to learn "factorized" representations of the task features. We then investigate the degree to which this factorization contributes to learning and generalization.

Methods

Networks We compared two simple forms of feedforward neural networks that differed only in how they computed stimulus similarity (Fig. 1). For the relational model, we implemented a simple form of the relational bottleneck (comparable to Kerg et al. (2022); Altabaa et al. (2023)), in which similarity was computed directly as the Euclidean distance between pairs of embeddings learned by the encoding layers; this was then used to determine the response (red pathway in Fig. 1). We compared this to a standard feedforward network, without any explicit similarity computation, in which the encoding layer projected to a multilayer perceptron (MLP) that generated the response (blue pathway in Fig. 1).

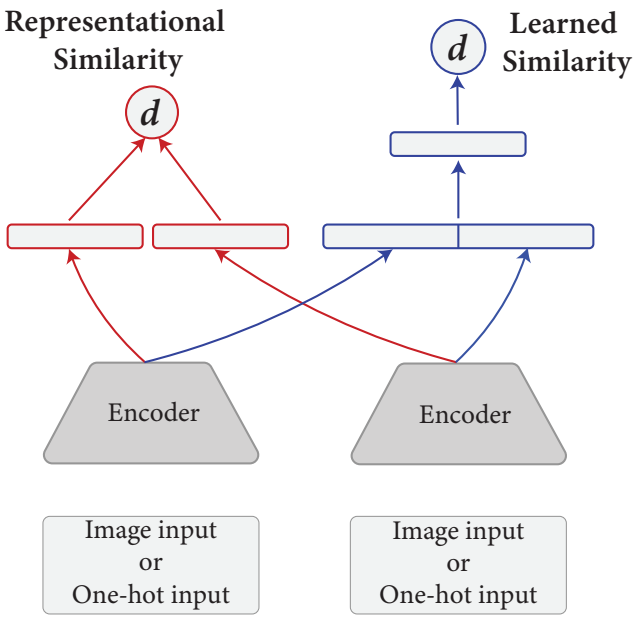

Figure 1: **Network Architecture** Networks trained with relational bottleneck (left) and learned similarity (right) used to perform similarity judgements between two input stimuli (see text).

Task and training data The task involved making similarity judgements over pairs of inputs that varied parametrically along two orthogonal dimensions. The inputs were pairs of grayscale images depicting geometric forms in which size and luminosity varied parametrically and independently across pairs (in the Appendix, we report similar results for the case in which features of the inputs were encoded as one-hots, rather than varying parametrically).

Results

First we examined both the rate at which the two networks learned to perform the task, and at which out of distribution (OOD) generalization improved — that is how quickly they learned to generalize performance to stimuli that were not only held out of the training corpus but that had feature values outside the range of those used in the training corpus. We found that that the relational network both learned to master the training data and exhibit OOD generalization substantially faster than the standard network (Fig. 2a). Next, we examined the structure of the representations learned in the encoding layers of the networks using PCA. Fig. 2b shows that the relational network learned orthogonal representations of the two feature dimensions, but that this was not the case for the standard network. These findings clearly indicate that imposing a simple form of the relational bottleneck in a standard feedforward network not only improves sample efficiency and the rate at which OOD generalization is achieved, but also imposes factorial structure on the representations learned immediately prior to the bottleneck — a form

of representation that is fundamental to compositional coding and that is known to be a feature of, and fundamental to the flexibility of human cognitive function. In the next set of simulations, we more directly examined the extent to which the relational bottleneck reproduces empirical observations about biases in human similarity judgments.

Dimensional Representations Align with Human Behavior

Here, we examined the extent to which the relational bottleneck can explain findings in human performance reflecting biases towards regularity (e.g., symmetry and parallelism) in geometric figures, and that have previously been hypothesized to reflect the presence of pre-specified symbolic representations (Sablé-Meyer et al., 2021). To do so, we trained the network described above (Figure 1a, red pathway) on an oddball detection task involving simple quadrilateral figures, used in Sablé-Meyer et al. (2021) to compare the performance of humans to non-human primates and artificial agents. That study was offered as evidence not only that humans exhibit a regularity bias in processing geometric figures that is not observed in non-human primates, but also that this bias can be captured by models that are explicitly imbued with pre-specified symbolic primitives but not standard deep learning neural networks that are trained directly on the images. Here, we tested whether a network imbued with a relational bottleneck — but not any symbolic primitives — could reproduce the regularity bias observed in humans.

Methods

The relational network described above (Fig. 1, red pathway) and SimCLR — a standard contrastive learning method, implemented here using a ResNet encoder — were trained on approximately $60,000$ trials using the same stimuli and following the same protocol used with humans and non-human primates in Sablé-Meyer et al. (2021). On each trial during the test phase, each network was presented with six images of quadrilateral figures, five of which were symmetric and identical in shape but varied in size and/or rotation, and one of which — the "oddball" — was different in shape (Fig. 3). The oddballs were constructed by perturbing the bottom right vertex of the reference shape to violate its regularity in the same way as in Sablé-Meyer et al. (2021). For each trial, we extracted the embedding in the model for each of the six images, and used these to identify the oddball as the one that was furthest from the centroid of the set (defined as the mean of the embeddings). We then computed the average error rate for each shape category and correlated the model error rates with the human and non-human primate error rates.

Results

We found that the network exhibited a pattern of performance as a function of both learning and regularity of shapes that closely resembled that observed by Sablé-Meyer et al. (2021) for humans but not for non-human primates (Fig. 4a-b). The network exhibited this pattern despite the fact that it was not endowed with any pre-specified symbolic primitives, nor any specific inputs that would have made these easier to discover. Rather, the results are attributable to the presence of the relational bottleneck, which has been shown in other settings to predispose to the discovery of low dimensional, abstract representations that can function as symbols (T. W. Webb et al., 2020; Kerg et al., 2022; Altabaa et al., 2023). Moreover, these biases towards human-like performance on this task emerge early during learning and are persistent over the course of network training (Fig. 4c). The relational training scheme yielded networks that learned well structured representations of the underlying category structure of the task (Fig 4d). While standard contrastive training has also been shown to produce networks sensitive to category information, our objective was to instead evaluate the degree of representational factorization within both relational and standard contrastive network architectures.

To do so, we assessed whether features associated with stimulus regularity were distinctly encoded in the network embeddings by applying linear regression models on the first 50 principal components of the networks' embeddings using 20-fold cross validation. The results indicated a more robust representation of stimulus regularity in the relational network ($R^2 = 0.89$) compared to the standard contrastive model ($R^2 = 0.68$). Notably, these findings were observed despite the fact that classifiers more accurately predicted the stimulus category from the standard contrastive networks' representations compared to the relational network representations (92% and 86%, respectively). This suggests that the improved decoding of stimulus regularity in the relational network is not due to a clearer representation of stimulus category, and is instead a product of the network's tendency to encode features along implicitly factorized and more linearly separable subspaces within the relational networks' embeddings.

Discussion

Low dimensional, factorized representations that can be used compositionally are generally thought to be a foundational requirement on which the flexibility of computation exhibited by humans is built. Most models that seek to describe such forms of computation have either used traditional symbolic processing mechanisms (Anderson, 2013; Newell, 1994) or have explicitly imbued neural networks with pre-specified symbolic primitives over which they can operate (Garcez et al., 2002). Some neural network models have used unsupervised learning coupled with specialized loss functions (Higgins et al., 2016; Kim & Mnih, 2018), but these require sensitive hyper-parameter tuning to discover dimensional structure. Another approach has been to supply the latent factors as fully supervised training targets with generative models (Tran et al., 2017). Our method provides an intermediate form of supervision between unsupervised and fully supervised methods. By using a scalar similarity target that implicitly captures the task's relational structure along the rele-

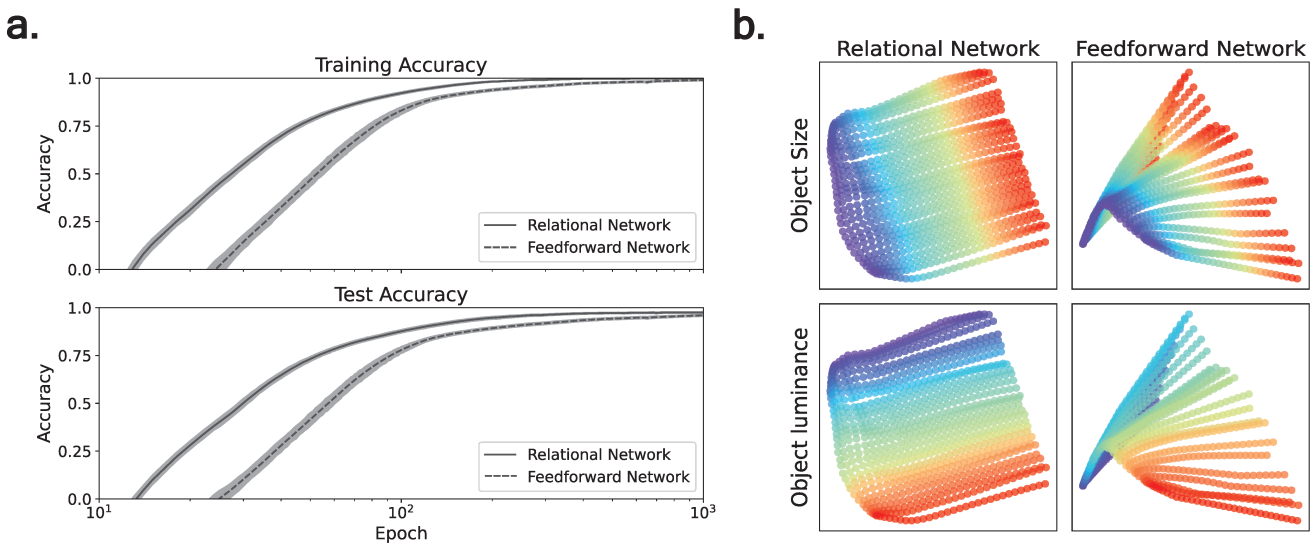

Figure 2: **(a)** Learning curves for training loss and generalization performance indicate that the relational architecture learns more rapidly than the feedforward architecture. **(b)** 2-dimension PCA of network embeddings learned by each network. Note that relational network learns orthogonal representations for each dimension, whereas the feed-forward network learns a non-linear manifold.

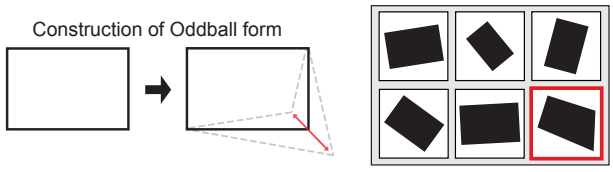

Figure 3: **Oddball construction & trial structure** Stimuli consisted of quadrilateral forms varying in their regularity/symmetry. Each trial was comprised of five variants of the same stimulus varying only in size and rotation, and one "oddball" stimulus constructed by perturbing the bottom right vertex of the reference stimulus to violate its regularity. Participants and networks were evaluated on their accuracy in identifying the oddballs.

vant dimensions, it implements a simple relational bottleneck (T. W. Webb et al., 2023) that, in turn, encourages the emergence of factorized representations, without explicitly providing them as targets. This approach affords the network the benefits of encoding task relevant features in its embeddings, while also remaining flexible enough to encode other features that may be relevant to the performance of other downstream tasks.

This method may also better account for how children construct factorized representations of their environments by receiving signals about the similarity structure of the world, either from cues in the environment or from explicit instruction by teachers and/or parents (Markman & Hutchinson, 1984; Gelman & Markman, 1986). Given the constancy of objects implicit in real-world environments, viewing these objects under multiple viewing conditions and in contrast to distinct objects may help carve out representations that factorize the relevant dimensions of variation. Furthermore, explicit instruction about the similarity of novel objects compared to known ones may implicitly provide information about the dimensional structure of the world that further facilitates factorization that can be exploited by architectures that include a form of relational bottleneck. In this way, agents may learn factorized representations of latent features without requiring explicit instruction about all of the particular features of the relevant stimuli. Instead, a training signal providing rich information about the similarity of objects in combination with the appropriate mechanisms for computing relations may be sufficient to leverage this comparatively weak form of supervision to learn factorized representations.

Furthermore, if an agent's experience does not encompass a sufficiently rich sampling of the underlying feature space, as is often the case in many naturalistic learning settings, relational architectures may insulate agents from the risks of overfitting that are common in more traditional architectures (Srivastava et al., 2014; Goodfellow et al., 2016). This protective effect is especially pronounced in environments where latent features are categorical, and stimuli consist of various combinations of these categorical features. In cases of sparse feature sampling, traditional neural networks tend to overfit by picking up on spurious correlations across independent feature dimensions. Relational bottleneck models, however, are more resistant to this, thereby reducing the likelihood of overfitting in such situations T. W. Webb et al., 2023. Moreover, the benefits of relational processing in promoting gen-

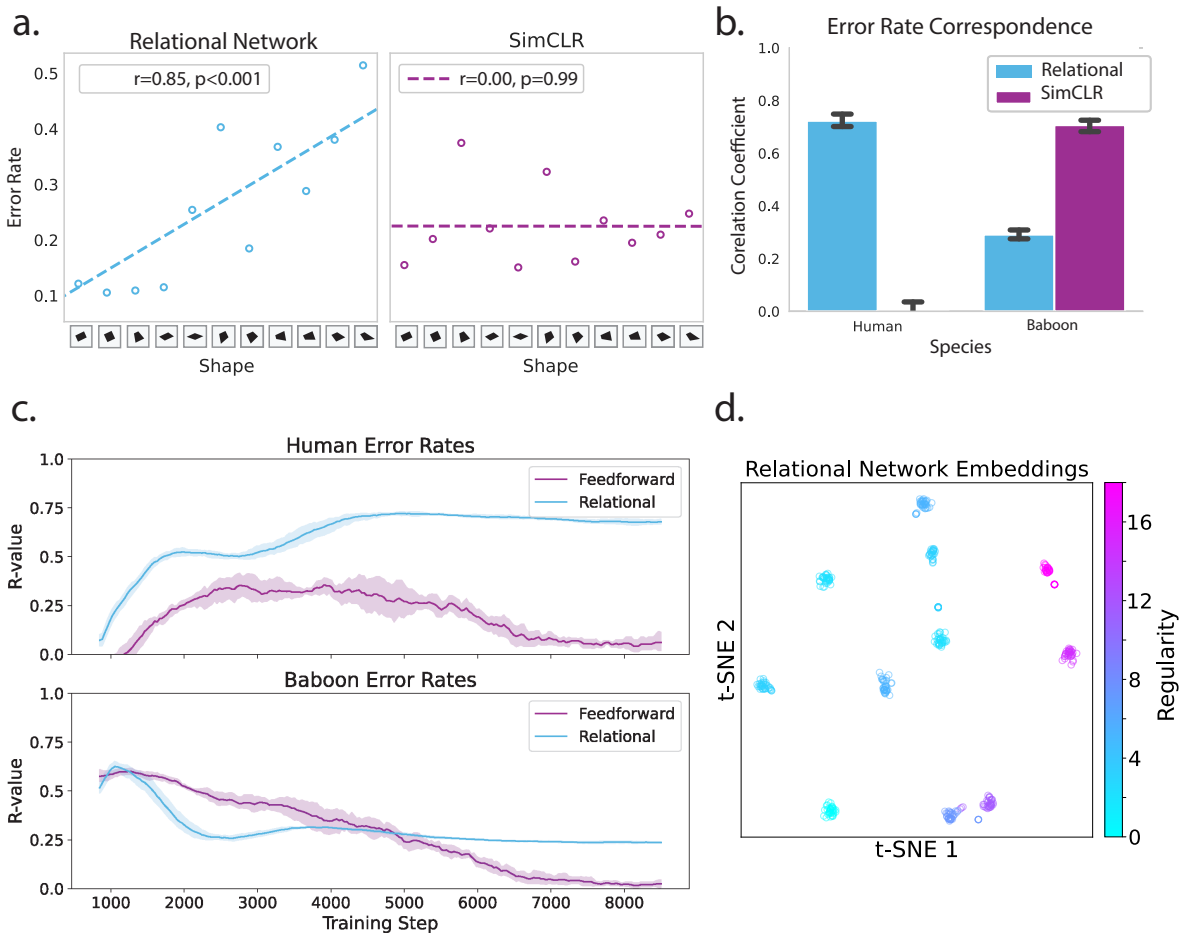

Figure 4: **Relational Network and SimCLR performance on oddball task: (a)** Error rates for the relational network and SimCLR at representative points during training. Note that the relational network's error rates exhibit a positive slope as a function of decreasing geometric regularity, consistent with human performance on this task, while the SimCLR network displays no sensitivity to geometric regularity. **(b)** Correlation coefficients between the model error rates and human and baboon error rates. Note that SimCLR most closely resembles baboon performance on this task while the relational network most closely corresponds with the human error rates. **(c)** t-SNE plot of reference shape embeddings from the relational network colored according to stimulus regularity calculated as the sum of the binary symbolic properties for each stimulus.

eralization and rapid learning are not only present in simple representation learning tasks as tested in the simulations reported here, but also across a range of challenging visual reasoning and analogy tasks (Mondal et al., 2023; T. W. Webb et al., 2020).

Previous work has demonstrated how augmenting neural networks with an explicit mechanism for computing relations provides substantial benefits in learning efficiency and generalization in navigational tasks (Whittington et al., 2020) and in basic reasoning tasks (Altabaa et al., 2023; T. W. Webb et al., 2020), approaching the sample efficiency and generalization abilities of human learners in these domains. Here, we show that these same computational elements may favor the formation of factorized upstream representations, that facilitate the discovery of relational structure.

This architecture may also provide structural/mechanistic insights into human brain function. Several studies have indicated that regions of the medial temporal lobe and hippocampus play an important role in navigation by providing a mechanism for binding features (Whittington et al., 2020) and computing the similarity of the current state with those stored in memory (Norman & O'Reilly, 2003; O'Reilly & McClelland, 1994). Such machinery for estimating similarity relations among distinct representations may provide a powerful architectural inductive bias not just useful memory retrieval, but also for factorizing representations. This hypothesis is consistent with recent evidence suggesting that regions that provide direct input to the hippocampus such as the entorhinal cortex encode highly factorized representations of space, time, and other cognitive variables (Fyhn et al., 2004;

Aronov et al., 2017; Constantinescu et al., 2016; Chandra et al., 2023). The work reported here provides one account for why these representations may emerge in regions providing input to the hippocampus. Furthermore, the framework suggests that other structures and mechanisms in the brain, that have similar functional attributes, may support relational abstraction and representational factorization in other domains (for example, the cerebellum and parietal cortex in the domain of motor function (Ravizza et al., 2006; D'Mello et al., 2020; McDougle et al., 2022)).

Conclusion

Our findings extend previous work, which has demonstrated that the use of a relational bottleneck (T. W. Webb et al., 2023) can induce a network to learn abstract rules and use these for extreme forms of generalization (T. W. Webb et al., 2020; Kerg et al., 2022; Altabaa et al., 2023). We show that the same inductive bias can induce the system to discover factorized, compositional representations of feature dimensions relevant to task performance in a data efficient manner, and in a form that approximates the efficiency of coding and flexibility of processing exhibited by the human brain.

References

Altabaa, A., Webb, T., Cohen, J., & Lafferty, J. (2023). Abstractors: Transformer modules for symbolic message passing and relational reasoning. *arXiv preprint arXiv:2304.00195*.

Anderson, J. R. (2013). *The adaptive character of thought*. Psychology Press.

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

Chandra, S., Sharma, S., Chaudhuri, R., & Fiete, I. (2023). High-capacity flexible hippocampal associative and episodic memory enabled by prestructured"spatial"representations. *bioRxiv*, 2023–11.

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*.

D'Mello, A. M., Gabrieli, J. D., & Nee, D. E. (2020). Evidence for hierarchical cognitive control in the human cerebellum. *Current Biology*, *30*(10), 1881–1892.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M.-B. (2004). Spatial representation in the entorhinal cortex. *Science*, *305*(5688), 1258–1264.

Garcez, A. S. d., Broda, K., & Gabbay, D. M. (2002). *Neural-symbolic learning systems: foundations and applications*. Springer Science & Business Media.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183–209.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations.*

Kerg, G., Mittal, S., Rolnick, D., Bengio, Y., Richards, B., & Lajoie, G. (2022). On neural architecture inductive biases for relational tasks. *arXiv preprint arXiv:2206.05056*.

Kim, H., & Mnih, A. (2018). Disentangling by factorising. In *International conference on machine learning* (pp. 2649–2658).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology*, *16*(1), 1–27.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. *MIT Press, Cambridge MA*, *3*, 44.

McDougle, S. D., Tsay, J. S., Pitt, B., King, M., Saban, W., Taylor, J. A., & Ivry, R. B. (2022). Continuous manipulation of mental representations is compromised in cerebellar degeneration. *Brain*, *145*(12), 4246–4263.

Mitchell, M., Palmarini, A. B., & Moskvichev, A. (2023). Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Mondal, S. S., Webb, T., & Cohen, J. D. (2023). Learning to reason over visual objects. *arXiv preprint arXiv:2303.02260*.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological review*, *110*(4), 611.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682.

Ravizza, S. M., McCormick, C. A., Schlerf, J. E., Justus, T., Ivry, R. B., & Fiez, J. A. (2006). Cerebellar damage produces selective deficits in verbal working memory. *Brain*, *129*(2), 306–320.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs.

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, *139*, 101527.

Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences*, *118*(16), e2023123118.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Tran, L., Yin, X., & Liu, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1415–1424).

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Webb, T. W., Frankland, S. M., Altabaa, A., Krishnamurthy, K., Campbell, D., Russin, J., ... Cohen, J. D. (2023). The relational bottleneck as an inductive bias for efficient abstraction. *arXiv preprint arXiv:2309.06629*.

Webb, T. W., Sinha, I., & Cohen, J. D. (2020). Emergent symbols through binding in external memory. *arXiv preprint arXiv:2012.14601*.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.