Relational reasoning and generalization using non-symbolic neural networks

Atticus Geiger[a,1], **Alexandra Carstensen**[b], **Michael C. Frank**[b], **and Christopher Potts**[a]

[a]Department of Linguistics, Stanford University; [b]Department of Psychology, Stanford University

This manuscript was compiled on December 13, 2020

Humans have a remarkable capacity to reason about abstract relational structures, an ability that may support some of the most impressive, human-unique cognitive feats. Because equality (identity) is a simple and ubiquitous relational operator, equality reasoning has been a key case study for the broader question of abstract relational reasoning. This paper revisits the question of whether equality can be learned by neural networks that do not encode explicit symbolic structure. Earlier work arrived at a negative answer to this question, but that result holds only for a particular class of hand-crafted feature representations. In our experiments, we assess out-of-sample generalization of equality using both arbitrary representations and representations that have been pretrained on separate tasks to imbue them with structure. In this setting, even simple neural networks are able to learn basic equality with relatively little training data. In a second case study, we show that sequential equality problems (learning ABA-patterned sequences) can be solved with only positive training instances. Finally, we consider a more complex, hierarchical equality problem, but this challenge requires vastly more data. However, using a pretrained equality network as a modular component of this larger task leads to good performance with no task-specific training ("zero-shot" generalization). Overall, these findings indicate that neural models are able to solve equality-based reasoning tasks, suggesting that essential aspects of symbolic reasoning can emerge from data-driven, non-symbolic learning processes.

relational reasoning | neural networks | generalization

One of the key components of human intelligence is our ability to reason about abstract relations between stimuli. Many of the most unremarkable human activities – scheduling a meeting, following traffic signs, assembling furniture – require a fluency with abstraction and relational reasoning that is unmatched in nonhuman animals. An influential perspective on human uniqueness holds that relational concepts are critical to higher-order cognition (e.g., 1). By far the most common case study of abstract relations has been equality.[*] Equality is a valuable case study because it is simple and ubiquitous, but also completely abstract in the sense that it can be evaluated regardless of the identity of the stimuli being judged.

Equality reasoning has been studied extensively across a host of systems and tasks, with wildly variant conclusions. In some studies, equality is very challenging to learn: only great apes with either extensive language experience or specialized training succeed in matching tasks in which a *same* pair, AA, must be matched to a novel same pair, BB (2, 3). Preschool children also struggle to learn these regularities in a seemingly similar task (4). In contrast, other studies suggest that equality is simple: bees are able to learn abstract identity relationships from only a small set of training trials (5, 6), and human infants can generalize identity patterns (7) and succeed in relational matching tasks (8). We take the central challenge of this literature to be characterizing the conditions that lead to success or failure in learning an abstract relation in a way that can be productively generalized to new stimuli (9).

The learning task in all of these cases can be described using the predicate *same* (or equivalently, $=$), which operates over two inputs and returns TRUE if they are identical in some respect. One perspective in the literature is that success in these learning tasks implies the presence of an equivalent symbolic description in the mind of the solver (2, 10). This view does not provide a lever to distinguish which of these tasks are trivial and which are difficult, however. Further, it can fall prey to circularity: because newborns show sensitivity to identity relations (11), then it would follow from this argument that they must have symbolic representations. If this logic applies also to bees, then we presuppose symbolic representations universally and have no account of the gradient difficulty of different tasks for different species.

An explanation of when same–different tasks are trivial and when they are difficult requires a theoretical framework beyond the symbolic/non-symbolic distinction. To make quantitative predictions about task performance, such a framework should ideally be instantiated in a computational model that takes in training data and learns a solution that generalizes when assessed with stimuli analogous to those used in experimental assessments. Symbolic computational models (e.g., 12) can be used to make contact with data about the breadth of generalization, but they require the existence of a symbolic equality predicate and hence again presuppose symbolic abili-

Significance Statement

The equality relation is a key case study for broader questions of abstract relational reasoning. Earlier work concluded that neural networks are incapable of learning general solutions to equality-based reasoning tasks, but this claim is in tension with the extraordinary recent success of neural networks in artificial intelligence. We resolve this tension by demonstrating these previous negative results hold only when neural networks are provided hand-crafted feature representations. When neural networks are instead provided the random or pretrained representations that are now the norm in state-of-the-art work, these models easily learn general solutions to equality-based reasoning tasks, suggesting that essential aspects of symbolic reasoning can emerge from non-symbolic learning processes.

All four authors contributed to paper writing and experimental design. Atticus Geiger and Chris Potts wrote the code that was used to run the experiments.

The authors declare no conflicts of interest.

[1] To whom correspondence should be addressed. E-mail: atticusg@stanford.edu

[*] We use the term "equality" here, though different literatures have also used "identity."

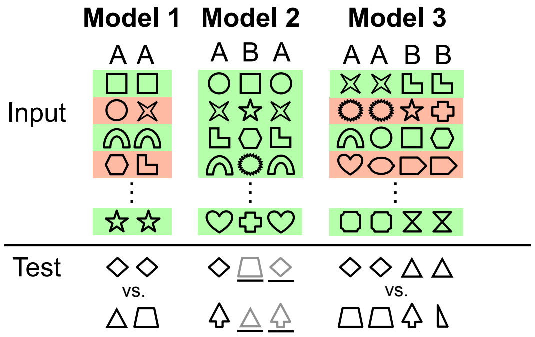


Fig. 1. Relational reasoning tasks. Green and red mark positive and negative training examples, respectively. The sequential task (Model 2) uses only positive instances, and a model succeeds if, prompted with α, it produces a sequence $\beta\ \alpha$ where $\beta \neq \alpha$. For the hierarchical task (Model 3), we show that a model trained on the basic task (Model 1) is effective with no additional training.

ties in every case of success. Ideally, we would want a model that describes under what conditions *same* is easy and under what conditions it is hard or unlearnable – and how learning proceeds in hard cases. Here, we aim to lay the foundation for the development of such an account.

We are inspired by an emergent perspective in the animal learning literature that the representations underlying non-human animals' and human infants' successes in equality reasoning tasks are graded (13). This view acknowledges the increasing evidence that other species like pigeons (14), crows (15), and baboons (16) can make true, out-of-sample generalizations of *same* and *different* relations, but it also recognizes that the observed patterns of behavior do not show the hallmarks of all-or-none symbolic representations. Instead, performance is graded. Out-of-sample generalization is possible but the level of performance depends critically on the diversity of the training stimuli (e.g., 17). Success requires hundreds, thousands, or even tens of thousands of training trials. And the outcome of learning is noisy and imperfect. These learning signatures appear to be a close match to the kind of learning exhibited by neural networks. Such networks are a flexible framework for arbitrary function learning, which have enjoyed a huge resurgence of interest in recent years in the fields of artificial intelligence, neuroscience, and cognitive science (e.g., 18, 19).

In an an influential rebuttal of the use of neural network models for capturing relational reasoning, Marcus et al. (10) argued that a broad class of recurrent neural networks were unable to learn equality relations. These claims were subsequently challenged by the presentation of evidence that some forms of neural networks are able to learn (at least aspects of) Marcus et al.'s equality tasks (20–24), yet these examples were not uncontroversial. The resulting debate (reviewed in 25) revealed a striking lack of consensus on some of the ground rules regarding what sort of generalization would be required to show that the learned function was suitably abstract.

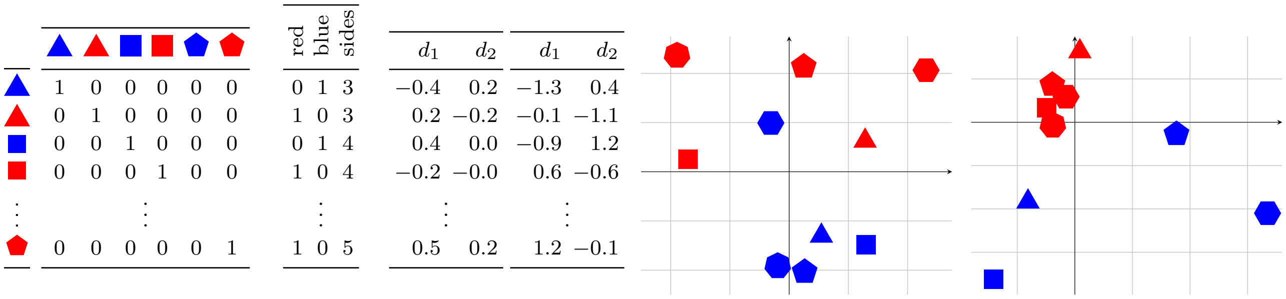
In the time since these debates, extraordinarily successful neural network models have been developed for tasks such as natural language inference (26, 27), question answering (28, 29), or visual reasoning (30), all of which are far more complex than equality-based tasks. In light of these findings, it may be surprising that the debate over equality-based reasoning is unresolved (25). Yet even recent work on equality-based reasoning tasks takes as its starting point the conclusion that neural networks are unable to succeed using standard architectures and general purpose learning algorithms (31–34). Further, though tasks and contexts vary, work in both computer vision (35–37) and machine reasoning (38–41) has presupposed that relational reasoning generally – and sometimes equality-based reasoning specifically – is difficult or impossible in standard network architectures.

Modern deep learning models have been so successful that it seems odd that they would be completely unable to learn equality-based reasoning tasks. We suspect these claims remain in the literature partly because only a narrow range of network architectures and representations were explored in the earlier debate, in part because it predated many important innovations in neural network design. Thus we revisit the debate here, using a broader range of architectures and representations and adopting stringent criteria for generalization. In particular we explore random and pretrained representations, which have facilitated some of the extraordinary successes of modern artificial intelligence (42–46). The use of pretrained representations to solve downstream tasks in particular is argued to be a hallmark of natural learning systems (47), has been an important feature of historical models from cognitive science (e.g., 48, 49), and is essential in the latest wave of state-of-the-art natural language processing models (46, 50–52).

In our current work, we model three cases of equality-based reasoning that have featured prominently in discussions of the role of symbols in relational reasoning (Figure 1): (1) learning to discriminate pairs of objects that exemplify the relation *same* or *different*, (2) learning sequences with repeated *same* elements (10), and (3) learning to distinguish hierarchical *same* and *different* relations in a context with pairs of pairs exemplifying these relations (2). Across these three models, we find strong support for their ability to learn equality relations. These results should serve to revise the conclusions of the earlier debate.

Marcus and colleagues (10, 53) showed experimentally that neural networks using feature representations cannot generalize to binary features unseen in training. We agree with this claim (and support it with a direct mathematical argument in SI Appendix). However, they concluded from this result that neural networks will need to have primitive symbolic operators to solve equality-based relational reasoning tasks, which is a solution that has been pursued in recent machine learning research (32–34). On this point, we disagree. Our experiments show that networks without such primitives can solve a range of these tasks using the sort of random or pretrained representations that are now the norm throughout artificial intelligence research. Overall, these findings suggest that some essential aspects of symbolic reasoning can emerge from entirely data-driven, non-symbolic learning processes.

Our work here makes three contributions. First, we resolve this longstanding debate by demonstrating neural networks are able to learn equality relations when provided with pretrained or random representations. Second, we modify the standard architecture of a recurrent neural network to allow it to learn the sequential equality task with no negative feedback. Negative evidence was dismissed as an unreasonably strong learning



	▲	▲	■	■	⬟	⬟	red	blue	sides	d_1	d_2	d_1	d_2
▲	1	0	0	0	0	0	0	1	3	−0.4	0.2	−1.3	0.4
▲	0	1	0	0	0	0	1	0	3	0.2	−0.2	−0.1	−1.1
■	0	0	1	0	0	0	0	1	4	0.4	0.0	−0.9	1.2
■	0	0	0	1	0	0	1	0	4	−0.2	−0.0	0.6	−0.6
⋮			⋮					⋮		⋮		⋮	
⬟	0	0	0	0	0	1	1	0	5	0.5	0.2	1.2	−0.1

(a) **(b)** Localist. **(c)** Property. **(d)** Random. **(e)** Pretrained. **(f)** The matrix in Figure 2d. **(g)** The matrix in Figure 2e.

Fig. 2. Each matrix is a method for representing the shapes in Figure 2a, where each row is a vector representation of one shape. Localist and property are featural representations where each vector dimension encodes the value of a single, semantically interpretable property. Random and pretrained are non-featural representations where the values of properties are encoded implicitly in two dimensions. Random representations and localist representations encode only identity, whereas property representations and pretrained representations encode color and number of sides.

regime in the original debate over these issues (e.g., 54), and we show that this learning regime is not necessary. Third, we show that a model pretrained on the simple equality task can achieve zero shot generalization to the hierarchical equality task, suggesting that pretraining might provide an account of how some organisms succeed on hard relational learning tasks. We believe these three contributions represent significant progress in our understanding of neural networks' ability to perform equality-based reasoning. Taken together, these contributions lay the groundwork for further non-symbolic neural network models of relational reasoning and abstract thought more broadly.

Designing theoretical models of equality learning

We begin by discussing two critical design considerations for our models: (1) the standards for generalization by which models should be evaluated and (2) the type of representations they should use. To summarize this discussion: we select generalization tasks with fully disjoint training and test vocabularies to provide the most stringent test of generalization. Further, we adopt both randomly initialized representations and pretrained representations for our subsequent models, and we show analytically that other kinds of representations are more limited in their capacity to make successful out-of-sample generalizations.

Generalization. The standard approach to training and evaluating neural networks is to choose a dataset, divide it randomly into training and assessment sets, train the system on the training set, and then use its performance on the assessment set as a proxy for its capacity to generalize to new data.

The standard approach is fine for many purposes, but it raises concerns in a context in which we are trying to determine whether a network has truly acquired a global solution to a target function. In particular, where there is any kind of overlap between the training and assessment vocabularies (primitive elements), we can't rule out that the network might be primarily taking advantage of idiosyncrasies in the underlying dataset to effectively cheat – to memorize aspects of the training set and learn a local approximation of the target function that happens to provide traction during assessment.

To address this issue, we follow (10) in proposing that networks must be evaluated on assessment sets that are completely disjoint in every respect from the train set, all the way down to the entities involved. For example, below, we train on pairs (a, a) and (a, b), where a and b are representations from a train vocabulary V_T. At test time, we create a new assessment vocabulary V_A, derive equality and inequality pairs (α, α) and (α, β) from that vocabulary, and assess the trained network on these new examples. In adopting these methods, we get a clear picture of the system's capacity to generalize, and we can safely say that its performance during assessment is a window into whether a global solution to identity has been learned. This is a very challenging setting for any machine learning model.

Representations. Essentially all modern machine learning models represent objects using vectors of real numbers. However, there are important differences in how these vectors are used to encode the properties of objects. The method of representation impacts whether there is a natural notion of similarity between entities and the ability of models to generalize to examples unseen in training. These two attributes are deeply related; if there is a natural notion of similarity between vector representations, then models can generalize to inputs with representations that are similar to those seen in training.

We characterize two broad approaches to such property encoding – which we call *featural representations* and *non-featural representations* – and argue that the differences between them have not been given sufficient attention in the debate about the ability of neural networks to perform relational reasoning. We acknowledge that a dimension of any vector representation is a "feature" but we adopt a usage that is common in cognitive science, namely that a feature is an interpretable semantic primitive.[†]

We ground our discussion in a hypothetical universe of blocks which vary by shape and color. Figure 2a is a partial view of them, and Figure 2b–Figure 2e present four different ways of encoding the properties of these objects in vectors.

Featural Representations. The defining characteristic of *featural* vector representations is that each dimension encodes the value

[†] The term *distributed representations* is used to refer jointly to what we call property representations, random representations, and pretrained representations. We opted not to use this term because it does not seperate property representations from random and pretrained representations, which is the relevant division here. Distributed representations are often contrasted in the neural network literature with *local* or *localist* representations; as discussed below, here we define these terms specifically to refer to representations whose features correspond to specific entities.

of a single, semantically intepretable property. The properties can be binary, integer-valued, or real-valued.

We use the term *localist* for the special case of featural representations in which only objects are represented and there is a feature corresponding to each object. In Figure 2b, each column represents the property of being an object, and every object is represented as a vector that has a single unit with value 1. There is no shared structure across objects; all are equally (un)related to each other as far as the model is concerned.

We will refer to featural representations that are not localist as *property representations*. Here, column dimensions encode specific, meaningful properties of objects. In our example, we can represent the properties of being red and being blue with two different binary features, and the property of having a certain number of sides as a single integer feature, as in Figure 2c. Unlike with localist representations, objects in this space can have complex relationships to each other, as encoded in the shared structure given by the columns.

Featural representations – both localist and property – have the appealing property that they are easy for researchers to interpret because of the tight correspondence between column dimensions and properties. However, this transparency actually inhibits neural networks from discovering general solutions. Instead, such models work far better with representations that have property values implicitly encoded in the abstract structure of the vector space. We demonstrate this result analytically in SI Appendixfor the case of binary features. The core insight is that networks cannot learn anything about column dimensions that are not represented in their training data; whatever weights are associated with those dimensions are unchanged by the learning process, so predictions about those dimensions remain random at test time. A developmental perspective suggests another reason to avoid binary featural representations, namely that this is not an accurate account of how perceptual inputs are represented in the brain.

Recent work in machine learning (32–34) attempts to overcome this analytic limitation of binary featural representations by modifying standard neural architectures to have symbolic primitives or changing network weight priors. In our work, we instead opt for non-featural representations, which do not have this analytic limitation and are the norm in state-of-the-art artificial intelligence models. There is no need to introduce symbolic primitives or modify network weight priors when non-featural representations are used.

Non-Featural Representations. A *non-featural representation* is a vector that encodes property values implicitly across many dimensions. Perhaps the simplest non-featural representations are *completely random* vectors, as in Figure 2d. Random representations can be seen as the non-featural counterpart to localist representations. In both of these representation schemes, all the objects are equally (un)related to each other, since column-wise patterns are unlikely in random representations and, to the extent that they are present, they exist completely by chance. However, in random representations, all the column dimensions can contribute meaningfully to identifying objects, whereas a localist representation has only one vector unit that determines the identity of any given object.

Random representations are a starting point that encodes object identity, but we can *pretrain* these representations via a learning process, imbuing them with rich structure that implicitly encodes property values across many dimensions. Figure 2e provides a simple example. This matrix is the results of pretraining the representations in Figure 2d on the task of predicting whether the object is blue, whether the object is red, and the number of sides the object has. (SI provides technical details on our pretraining approach.) Superficially, the two matrices look equally random, but the random representations in Figure 2f have no such structure, while the pretrained representations in Figure 2g do: there is a line that separates blue and red objects.

Pretraining need not be restricted to input representations; all the parameters of a model can be pretrained, offering the possibility that networks might be used as modular components to solve more complex tasks. We realize this possibility with our third experiment, where a model pretrained on a simple equality is used as a modular component to compute hierarchical equality.

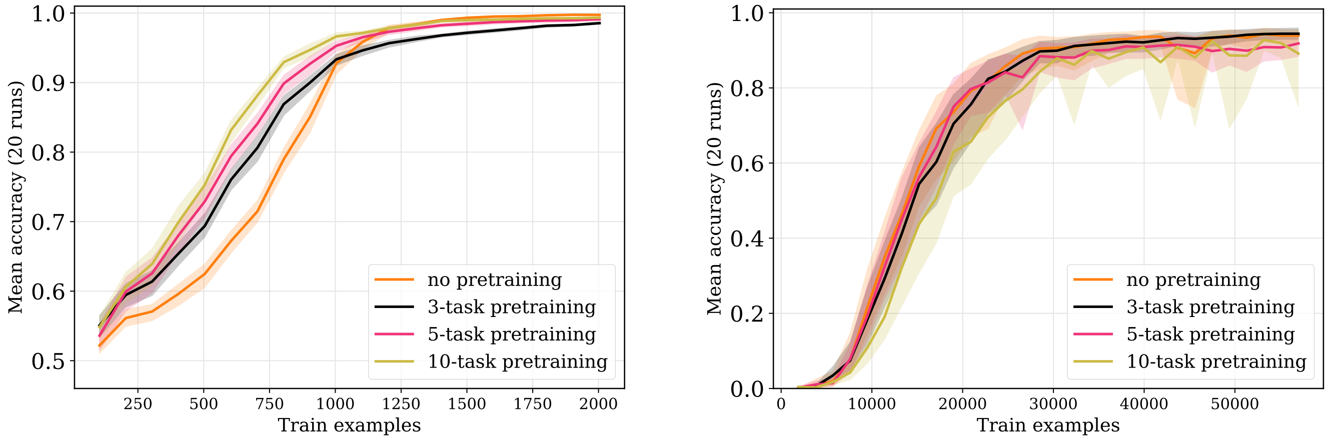
Model 1: Same–different relations

First, we investigate whether a supervised single layer feed-forward neural network can learn the equality relation in the strict setting we describe above where train and test vocabulary are disjoint. The input is a pair of vectors (a, b) which correspond to the two stimulus objects. These vectors are non-featural representations that do not have features encoding properties of the objects or their identity. During training, this model is presented with positive and negative labeled examples. During testing, this model is tasked with categorizing inputs unseen during training. It is straightforward to show that a network like this is capable of learning equality as we have defined it. Indeed, in our SI Appendix, we provide an analytic solution to the equality relation using this neural model.

This result shows that equality in our sense is learnable in principle, but it doesn't resolve the question of whether networks can find this kind of solution given finite training data. To address this issue, we train networks on a stream of pairs of random vectors. Half of these are identity pairs (a, a), labeled with 1, and half are non-identity pairs (a, b), labeled with 0. Trained networks are assessed on the same kind of balanced dataset, with vectors that were never seen in training so that, as discussed earlier, we get a clear picture of whether they have found a generalizable solution.

Results. Figure 3a shows our results. The representations used in these experiments are random representations that were pretrained using a linear classifier for 0, 3, 5, or 10 different binary feature discrimination tasks. For example, following Figure 2, a three-task model might be trained to encode the binary properties of being blue, having four sides, and being red. For all representations, this neural model reached above-chance performance almost immediately, but required upwards of 1,000 examples to achieve near perfect accuracy. Interestingly, we observed a clear speed-up, with more pretraining tasks resulting in the largest gains. It seems that, by grounding our representations in "property domains" (as represented by the different task dimensions), we imbue them with implicit structure that makes learning easier.

Discussion. Our assessment pairs have nothing in common with the training pairs except insofar as both involve vectors of



(a) Results for single layer feed-forward neural networks trained on our simple equality task. The 'no pretraining' model is provided random representations and the 'k-task pretraining' models are provided random representations are grounded in k binary property domains via pretraining learning tasks.

(b) Results for LSTM recursive neural networks trained on our sequential equality task. The 'no pretraining' model is provided random representations and the 'k-task pretraining' models are provided random representations are grounded in k binary property domains via pretraining learning tasks. All the training examples are presented at once over multiple epochs

Fig. 3. Results for (a) the simple equality task and (b) the sequential equality task.

real numbers of the same dimensionality. During training, the network is told (via labels) which pairs are equality pairs and which are not, but the pairs themselves contain no information about equality per se. It thus seems fair to us to say that these networks have learned equality – or at least how to simulate that relation with near perfect accuracy. Further, the use of representations that are structured by pretraining results in faster learning.

Model 2: Sequential same–different (ABA task)

Our first model is simple and successfully learns equality. However, this model is supervised with both positive and negative evidence. In the initial debate around these issues, supervision with negative evidence was dismissed as an unreasonably strong learning regime (e.g., 54). While this argument likely holds true for language learning (in which supervision is generally agreed not to be binary or direct; 55, 56), it is not necessarily true for learning more generally. Nevertheless, learning of sequential rules without negative feedback is possible for infants (10, 57). In experiments of this type, infants are presented with a set of positive examples. Our next model explores whether neural network models can learn this task in a challenging regime with no negative supervision.

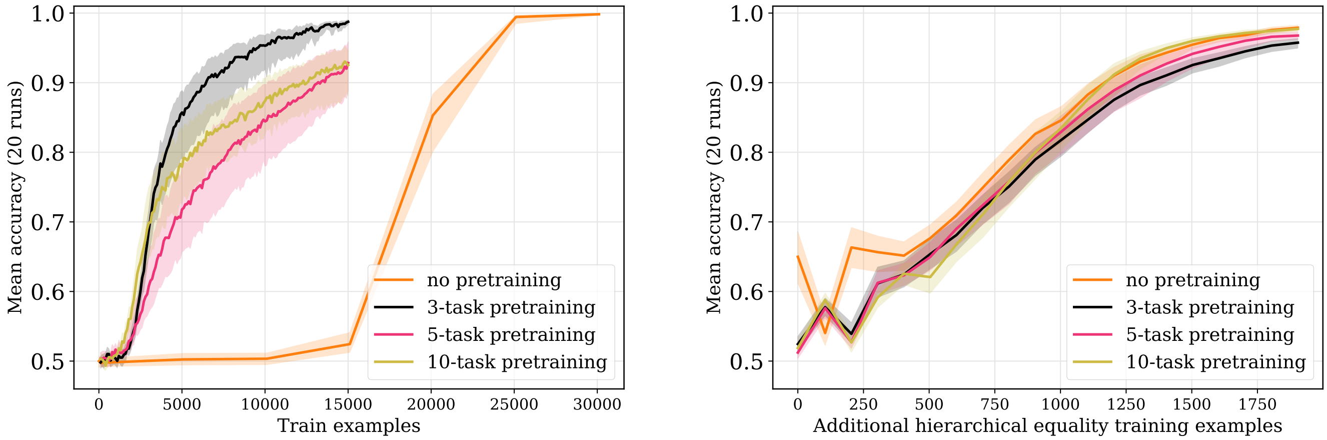
To explore learning with only positive instances, we use a neural LSTM language model, a recursive network with the ability to selectively forget and remember information (58). Language models are sequential: at each timestep, they predict an output given their predictions about the preceding timesteps. As typically formulated, the prediction function is just a classifier: at each timestep, it predicts a probability distribution over the entire vocabulary of options, and the item with the highest probability is chosen as a symbolic output. This output becomes the input at the next timestep, and the process continues.

This formulation will not work in situations in which we want to make predictions about test items with an entirely disjoint vocabulary from the training sample. The classifier function will get no feedback about these out-of-vocabulary items during training, and so it will never predict them during testing. To address this issue, we reformulate the prediction function. Our proposal is to have the model predict output vector representations – instead of discrete vocabulary items – at each timestep. During training, the model is trained to minimize the distance between these output predictions and the representations of the actual output entities. During assessment, we take the prediction to be the item in the entire vocabulary (training and assessment) whose representation is closest to the predicted vector (in terms of Euclidean distance). This fuzzy approach to prediction creates enough space for the model to predict sequences from an entirely new vocabulary. Our SI Appendixprovide an analytic solution to the ABA task using this model.

To see how well the model performs in practice, we trained networks on sequences `<s> a b a </s>`, where $\texttt{b} \neq \texttt{a}$. We show the network every such sequence during training, from an underlying vocabulary of 20 items (creating a total of 380 examples). To assess how well the model learns this pattern, we seed it with `<s> x` where `x` is an item from a disjoint vocabulary from that seen in training, and we say that a prediction is accurate if the model continues with `y x </s>`, where `y` is any character (from the training or assessment vocabulary) except `x`.

Results. Figure 3b shows our results. Unlike for the previous equality experiment, we found that we had to allow the model to experience multiple epochs of training on the same set in order to succeed and tens of thousands of training examples were necessary. We considered a range of representations (as in Model 1); the model was again successful with all representations, but in this experiment pretraining representations did not increase performance.

Discussion. These sequential models are given no negative examples and they must predict into a totally new vocabulary. Despite these challenges, they succeed at learning the



(a) Results for two layer feed-forward neural networks trained on our hierarchical equality task. The 'no pretraining' model is provided random representations and the 'k-task pretraining' models are provided random representations are grounded in k binary property domains via pretraining learning tasks.

(b) Results for simple equality networks applied to the hierarchical equality task. The 'no pretraining' model is provided random representations and the 'k-task pretraining' models are provided random representations are grounded in k binary property domains via pretraining learning tasks.

Fig. 4. Results for(a) hierachical sequential equality task without pretraining on simple equality and (b) with pretraining on simple equality.

underlying patterns in our data. On the other hand, the learning process is slow and data-intensive. We hypothesized that grounding representations in property domains via pretraining might lead to noticeable speed-ups, as it did in for our simple same-different task, but we did not see this effect in practice. We speculate that there may be model variants that reduce these demands, given that learning is in principle possible in this architecture, but we leave them to future work.

Model 3: Hierarchical same–different relations

Given the strong results found for simple equality relations, we can ask whether more challenging equality problems are also learnable in our setting. The hierarchical equality task used by Premack (2) is an interesting test case: given a pair of pairs $((a, b), (c, d))$, the label is 1 if $(a = b) = (c = d)$, else 0. Premack suggested that the ability exemplified by this task – reasoning about hierarchical *same* and *different* relations – could represent a form of symbolic abstraction uniquely enabled by language. Given the non-symbolic nature of our models, our simulations provide a test of this hypothesis, though we should look critically at their ability to find good solutions with reasonable amounts of training data.

We can approach this task using the same model and methods as we used for equality, with the relatively minor change of providing the network four vector representations instead of two. We found that single layered feed-forward neural networks required nearly 100,000 training examples to solve this task. We hypothesized that a single layer network might be suboptimal here. This task is intuitively hierarchical: if one works out the equality labels for each of the two pairs, then the further classification decision can be done entirely on that basis. Our current neural network might be too shallow to find this kind of decomposition. To address this issue, we use a two layer feed forward network.

Results without pretraining. Figure 4a shows our results. We again considered a range of representations and again the network succeeded across this range, with pretraining increasing performance as in Model 1. The network required more than 20,000 training instances to reach top performance, and upwards of 10,000 examples with pretrained representations.

This amount of training data is vastly more data than human participants get in similar experiments, which typically involve short exposures in the range of dozens to hundreds of examples (e.g., 10, 59). Thus, it is worth asking whether there are other solutions that would be more data efficient and more in line with human capabilities. We next seek to further capitalize on the hierarchical nature of this task by defining a modular pretraining regime in which previously learned capabilities are recruited for new tasks.

The critical role of experience. Our successful results training neural networks on simple equality suggested another strategy for solving the hierarchical equality task. Rather than requiring our networks to find solutions from scratch, we pretrained them on basic equality tasks and then used those parameters as a starting point for learning hierarchical equality. This set of simulations was conceptually similar to our previous experiments with pretrained input representations, but now we pretrained an entire subpart of the model, rather than just input representations. This approach parallels the experimental paradigm used by Thompson et al. (60), in which chimpanzees that received pre-training on a basic equality (same/different judgment) task – but not naive champanzees – succeed in a hierarchical equality task.

The hierarchical equality task requires computing the equality relation three times: compute whether each pair of inputs are equal and then compute whether the truth-valued outputs of these first two computations are equal. We thus used the same network pretrained on basic equality to perform all three equality computations.

Figure 4b shows our results. All the models have above chance performance after being trained only on the simple equality task – that is, they achieve zero-shot generalization to the hierarchical task and within two thousand examples, the models achieve near perfect accuracy. It is remarkable that a model trained only on equality between entities is able to get traction on a problem that requires determining whether equality holds between the truth values encoded in two learned representations. Pretrained representations did not increase performance.

General Discussion

Equality is a key case study for understanding the origins of human relational reasoning. This case study has been puzzling for symbolic accounts of reasoning because such accounts do not provide a compelling explanation for why some equality tasks are so easy to learn and others are so hard. In addition, evidence of graded learning and generalization in non-human species suggests that a gradual learning account might provide more traction in explaining the empirical data (13). Inspired by this work, we revisited a long-standing debate about whether neural network models can learn equality relations from data (25).

Our work here makes three contributions to this debate. First, we show that non-featural representations — both random and pretrained – allow standard neural networks to learn simple, sequential, and hierarchical equality tasks. Both the research that originated this debate (10, 20–24, 53) and more recent work (31–34) only involve experiments where featural representations are used, and we suggest that this choice led directly to the conclusions of this body of work. Second, we show that neural networks can achieve high test accuracy on the sequential equality task with no negative feedback, suggesting that a negative feedback learning regime is not critical for learning equality. Finally, we show that a neural network trained only on simple equality can generalize to hierarchical equality, even in a "zero-shot" evaluation. Although pretrained representations sometimes led to faster learning, they were not a necessary component for models to succeed, and success was possible even using random representations.

In some settings, our current models require many more training instances than humans seem to need. However, our pretraining approach suggests a path forward: by using pretrained models as modular components, we can get traction on challenging tasks without any training specifically for those tasks. In some cases, even a small amount of additional training can make a substantial difference. Perhaps pretrained components of this type could serve as the basis for more complex cognitive abilities more generally (61, 62). Other computational work on relational reasoning in cognitive science has developed hybrid architectures that do not explicitly encode symbolic equality but incorporate other symbolic structures (e.g., 63); integrating functions learned from data could be an interesting direction for future work with these architectures. Similarly, other work on relational reasoning in artificial intelligence has used networks with explicitly relational architectures – such architectures could be interesting to explore in the context of equality reasoning (38–41).

One further implication of our pretraining findings is that it should be possible to scaffold non-human animals' performance in complex, hierarchical equality tasks via training on simpler ones. Indeed, (15) show just this result in crows, consistent with our findings. Although we do not discount the potential role of linguistic labels in informing adult humans' expertise in such tasks (64), pretraining also provides a potential account of how infants and young children might succeed in a range of equality reasoning tasks without access to specific linguistic symbols like "same" (4, 8, 65).

More broadly still, our work suggests a possible way forward in understanding the acquisition of logical semantics. Graded logical functions like those our models learned here could form the foundation for a semantics of words like "same" (66). Such an option is appealing because it escapes from the circularity of defining the semantics of linguistic symbols as originating in a mental primitive SAME. A semantics for "same" requires defining its inputs and outputs as well as how it composes with other symbols. The assertion that there is a primitive identity computation does not specify the format of these inputs and outputs or these composition rules; it further fails to explain the flexibility that allows us to call two Toyota Corollas "the same" but two twin sisters "different." In contrast, the kinds of networks we propose here could in principle be conditioned contextually to provide flexible, context-sensitive interpretation of logical meaning. Such a contextually-conditioned semantics could be applied in both cases of sameness and graded similarity (67), holding the promise of unifying models of identity and similarity.

Earlier debates about the nature of equality computations centered around the question of whether models included symbolic elements. We believe ours do not; but it is of course possible to quibble with this judgment. For example, since the supervisory signal used in Models 1 and 3 is generated based on a symbolic rule, perhaps that makes these models symbolic under some definition. (Of course, the same argument could be applied to the supervision signal that is provided to crows, baboons, and human children in some tasks). We view this kind of argument as terminological, rather than substantive. In the end, our goal is an explicit learning theory for relational reasoning. Our hope is that the work described here takes a first step in this direction.

Materials and Methods

The code and data used to run the experiments in this paper are publicly available at: https://github.com/atticusg/NeuralRelationalReasoning

The simple equality model takes in input vectors, a, b, and uses a single linear transformation followed by a non-linearity to create a hidden representation h that is then used to create a probability distribution, y, over the two classes.

$$h = \mathrm{ReLU}([a; b]W_{xh} + b_h) \qquad y = \mathbf{softmax}(hW_{hy} + b_y) \qquad [1]$$

The sequential equality model takes in a sequence of input vectors, $x_1, x_2, \ldots$, and uses an LSTM cell to create a hidden representation h_t at each timestep t that is that is then linearly projected into the input vector space providing a prediction for that timestep, y_t.

$$h_t = \mathbf{LSTM}(x_t, h_{t-1}) \qquad y_t = h_t W + b \qquad [2]$$

The first hierarchical equality model takes in input vectors, a, b, c, d, and applies a linear transformation followed by a non-linearity to create a hidden representation h_1 and then applies these two steps once more to create second hidden representation h_2 that is then used to create a probability distribution, y, over the two classes.

$$h_1 = \mathrm{ReLU}([a; b; c; d]W_{xh} + b_{h_1}) \qquad [3]$$

$$h_2 = \mathrm{ReLU}(h_1 W_{hh} + b_{h_2}) \qquad y = \mathbf{softmax}(h_2 W_{hy} + b_y) \qquad [4]$$

The second hierarchical equality model takes in input vectors, a, b, c, d, and applies the simple equality model from equations (1) to the pairs (a, b) and (c, d) to produce hidden representations h_1 and h_2. Then the simple equality model is applied once again to the pair (h_1, h_2) to produce a final hidden representation h_3 that is used to create a probability distribution, y, over the two classes.

$$h_1 = \mathrm{ReLU}([a; b]W_{xh} + b_h) \qquad h_2 = \mathrm{ReLU}([c; d]W_{xh} + b_h) \quad [5]$$

$$h_3 = \mathrm{ReLU}([h_1; h_2]W_{xh} + b_h) \qquad y = \mathbf{softmax}(h_3 W_{hy} + b_y) \quad [6]$$

The parameters for the simple and hierarchical equality models are learned using back propagation with a cross entropy objective function defined as follows, for a set of N examples and K classes:

$$\max(\theta) \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y^{i,k}\log(h_\theta(i)^k) \qquad [7]$$

where θ abbreviates the model parameters, $y^{i,k}$ is the actual label for example i and class k, and $h_\theta(i)^k$ is the corresponding prediction.

The parameters for the sequential model are learned using back propagation with a squared mean error objective function defined as follows:

$$\max(\theta) \quad -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\left\| h_\theta\left(x^{i,0:t-1}\right) - x^{i,t}\right\|^2 \qquad [8]$$

for N examples. Here, T_i is the length of example i. As before, θ abbreviates the parameters of the model. We use $h_\theta(x^{i,0:t-1})$ for the vector predicted by the model for example i at timestep t, which is compared to the actual vector at timestep t via squared Euclidean distance.

Hyperparameter searches and implementation details can be found in SI.

ACKNOWLEDGMENTS. This work is supported in part by a Facebook Robust Deep Learning for Natural Language Processing Research Award and by the McDonnell Foundation grant "The Nature and Origins of the Human Capacity for Abstract Combinatorial Thought".

1. Gentner D, Goldin-Meadow S (2003) *Language in mind: Advances in the study of language and thought*. (MIT press).
2. Premack D (1983) The codes of man and beasts. *Behavioral and Brain Sciences* 6(1):125–136.
3. KR Thompson R, Rattermann MJ, L Oden D (2001) Perception and judgement of abstract same-different relations by monkeys, apes and children: Do symbols make explicit only that which is implicit? *Hrvatska revija za rehabilitacijska istraživanja* 37(1):9–22.
4. Walker CM, Bridgers S, Gopnik A (2016) The early emergence and puzzling decline of relational reasoning: Effects of knowledge and search on inferring abstract concepts. *Cognition* 156:30–40.
5. Giurfa M, Zhang S, Jenett A, Menzel R, Srinivasan MV (2001) The concepts of 'sameness' and 'difference' in an insect. *Nature* 410(6831):930–933.
6. Avarguès-Weber A, Deisig N, Giurfa M (2011) Visual cognition in social insects. *Annual review of entomology* 56:423–443.
7. Anderson EM, Chang YJ, Hespos S, Gentner D (2018) Comparison within pairs promotes analogical abstraction in three-month-olds. *Cognition* 176:74–86.
8. Ferry AL, Hespos SJ, Gentner D (2015) Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development* 86(5):1386–1405.
9. Carstensen A, Frank MC (in press) Do graded representations support abstract thought? *Current Opinion in Behavioral Sciences*.
10. Marcus GF, Vijayan S, Rao SB, Vishton PM (1999) Rule learning by seven-month-old infants. *Science* 283(5398):77–80.
11. Gervain J, Berent I, Werker JF (2012) Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience* 24(3):564–574.
12. Frank MC, Tenenbaum JB (2011) Three ideal observer models for rule learning in simple languages. *Cognition* 120(3):360–371.
13. Wasserman E, Castro L, Fagot J (2017) Relational thinking in animals and humans: From percepts to concepts. in *APA HAndbook of Comparative Psychology: Perception, Learning, and Cognition*, eds. Call J, Burghardt GM, Pepperberg IM, Snowdon CT, Zentall T. (American Psychological Association) Vol. 2.
14. Cook RG, Wasserman EA (2007) Learning and transfer of relational matching-to-sample by pigeons. *Psychonomic Bulletin & Review* 14(6):1107–1114.
15. Smirnova A, Zorina Z, Obozova T, Wasserman E (2015) Crows spontaneously exhibit analogical reasoning. *Current Biology* 25(2):256–260.
16. Fagot J, Thompson RK (2011) Generalized relational matching by guinea baboons (papio papio) in two-by-two-item analogy problems. *Psychological Science* 22(10):1304–1309.
17. Castro L, Kennedy PL, Wasserman EA (2010) Conditional same-different discrimination by pigeons: Acquisition and generalization to novel and few-item displays. *Journal of Experimental Psychology: Animal Behavior Processes* 36(1):23.
18. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *nature* 521(7553):436–444.
19. Saxe AM, McClelland JL, Ganguli S (2019) A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* 116(23):11537–11546.
20. Dienes Z, Altmann GT, Gao SJ (1999) Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science* 23(1):53–82.
21. Seidenberg MS, Elman JL (1999) Networks are not 'hidden rules'. *Trends in Cognitive Sciences* 3(8):288–289.
22. Seidenberg MS, Elman JL (1999) Do infants learn grammar with algebra or statistics? *Science* 284(5413):433–433.
23. Elman JL (1999) Generalization, rules, and neural networks: A simulation of Marcus et. al. HTML document.
24. Negishi M (1999) Do infants learn grammar with algebra or statistics? *Science* 284(5413):435.
25. Alhama RG, Zuidema W (2019) A review of computational models of basic rule learning: The neural-symbolic debate and beyond. *Psychonomic bulletin & review* 26(4):1174–1194.
26. Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (Association for Computational Linguistics).
27. Williams A, Nangia N, Bowman S (2018) A broad-coverage challenge corpus for sentence understanding through inference in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. (Association for Computational Linguistics), pp. 1112–1122.
28. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. (Association for Computational Linguistics, Austin, Texas), pp. 2383–2392.
29. Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: Unanswerable questions for SQuAD in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. (Association for Computational Linguistics, Melbourne, Australia), pp. 784–789.
30. Johnson J, et al. (2017) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1988–1997.
31. Alhama RG, Zuidema W (2018) Pre-wiring and pre-training: What does a neural network need to learn truly general identity rules? *Journal of Artificial Intelligence Research* 61:927–946.
32. Weyde T, Kopparti R (2019) Modelling identity rules with neural networks. *Journal of Applied Logic* 6(4):745–769.
33. Weyde T, Kopparti R (2018) Feed-forward neural networks need inductive bias to learn equality relations in *Proceedings of the NeurIPS 2018 Relation Representation Learning Workshop*.
34. Kopparti R, Weyde T (2020) Weight priors for learning identity relations.
35. Kim J, Ricci M, Serre T (2018) Not-so-CLEVR: learning same–different relations strains feedforward neural networks. *Interface Focus* 8.
36. Fleuret F, et al. (2011) Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences of the United States of America* 108:17621–5.
37. Gülçehre c, Bengio Y (2016) Knowledge matters: Importance of prior information for optimization. *Journal of Machine Learning Research* 17(1):226–257.
38. Raposo D, et al. (2017) Discovering objects and their relations from entangled scene representations. *arXiv preprint arXiv:1702.05068*.
39. Santoro A, et al. (2017) A simple neural network module for relational reasoning in *Advances in neural information processing systems*. pp. 4967–4976.
40. Santoro A, et al. (2018) Relational recurrent neural networks in *Advances in neural information processing systems*. pp. 7299–7310.
41. Palm R, Paquet U, Winther O (2018) Recurrent relational networks in *Advances in Neural Information Processing Systems*. pp. 3368–3378.
42. Collobert R, et al. (2011) Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537.
43. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality in *Advances in Neural Information Processing Systems 26*, eds. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ. (Curran Associates, Inc.), pp. 3111–3119.
44. Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (Association for Computational Linguistics, Doha, Qatar), pp. 1532–1543.
45. Peters M, et al. (2018) Deep contextualized word representations in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. (Association for Computational Linguistics), pp. 2227–2237.
46. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. (Association for Computational Linguistics, Minneapolis, Minnesota), pp. 4171–4186.
47. Zador AM (2019) A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications* 10(1):1–7.
48. Landauer TK, Dumais ST (1997) A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2):211.
49. McRae K, Hetherington PA (1993) Catastrophic interference is eliminated in pretrained networks in *Proceedings of the 15h annual conference of the cognitive science society*. pp. 723–728.
50. Liu Y, et al. (2019) Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692.
51. Radford A, et al. (2019) Language models are unsupervised multitask learners. *OpenAI Blog*.
52. Brown T, et al. (2020) Language models are few-shot learners. *ArXiv* abs/2005.14165.
53. Marcus GF (2001) *The Algebraic Mind: Integrating Connectionism and Cognitive science*.

(MIT Press).

54. Marcus GF (1999) Rule learning by seven-month-old infants and neural networks. Response to Altmann and Dienes. *Science* 284:875.
55. Brown R, Hanlon C (1970) Derivational complexity and order of development in speech in *Cognition and the development of language.*, ed. Hayes JR. (Wiley).
56. Chouinard MM, Clark EV (2003) Adult reformulations of child errors as negative evidence. *Journal of child language* 30(3):637–669.
57. Rabagliati H, Ferguson B, Lew-Williams C (2019) The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental science* 22(1):e12704.
58. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural computation* 9(8):1735–1780.
59. Endress AD, Scholl BJ, Mehler J (2005) The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General* 134(3):406.
60. Thompson RKR, Oden DL, Boysen ST (1997) Language-naive chimpanzees (pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes* 23(1):31—-43.
61. Frank MC, Everett DL, Fedorenko E, Gibson E (2008) Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition* 108(3):819–824.
62. Heyes C (2018) *Cognitive gadgets: The cultural evolution of thinking.* (Harvard University Press).
63. Hummel JE, Holyoak KJ (2003) A symbolic-connectionist theory of relational inference and generalization. *Psychological review* 110(2):220.
64. Gentner D (2003) Why we're so smart. *Language in mind: Advances in the study of language and thought* 195235.
65. Hochmann JR, Mody S, Carey S (2016) Infants' representations of same and different in match-and non-match-to-sample. *Cognitive psychology* 86:87–111.
66. Potts C (2019) A case for deep learning in semantics: Response to pater. *Language.*
67. Medin DL, Goldstone RL, Gentner D (1993) Respects for similarity. *Psychological review* 100(2):254.