Connection Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ccos20

An emergent approach to analogical inference

Paul H. Thibodeau [a] , Stephen J. Flusberg [b] , Jeremy J. Glick [c] & Daniel A. Sternberg [d]

[a] Department of Psychology , Oberlin College , Oberlin , OH , 44074 , USA

[b] Purchase College, State University of New York , Purchase , NY , USA

[c] Disney Interactive Media Group , Palo Alto , CA , USA

[d] Lumos Labs, Inc. , San Francisco , CA , USA

To cite this article: Paul H. Thibodeau , Stephen J. Flusberg , Jeremy J. Glick & Daniel A. Sternberg (2013) An emergent approach to analogical inference, Connection Science, 25:1, 27-53, DOI: 10.1080/09540091.2013.821458

To link to this article: http://dx.doi.org/10.1080/09540091.2013.821458

Connection Science, 2013
Vol. 25, No. 1, 27–53, http://dx.doi.org/10.1080/09540091.2013.821458

Taylor & Francis
Taylor & Francis Group

An emergent approach to analogical inference

Paul H. Thibodeau[a]*, Stephen J. Flusberg[b], Jeremy J. Glick[c] and Daniel A. Sternberg[d]

[a]*Department of Psychology, Oberlin College, Oberlin, OH 44074, USA;* [b]*Purchase College, State University of New York, Purchase, NY, USA;* [c]*Disney Interactive Media Group, Palo Alto, CA, USA;* [d]*Lumos Labs, Inc., San Francisco, CA, USA*

(*Received 15 August 2012; final version received 28 June 2013*)

In recent years, a growing number of researchers have proposed that analogy is a core component of human cognition. According to the dominant theoretical viewpoint, analogical reasoning requires a specific suite of cognitive machinery, including explicitly coded symbolic representations and a mapping or binding mechanism that operates over these representations. Here we offer an alternative approach: we find that analogical inference can emerge naturally and spontaneously from a relatively simple, error-driven learning mechanism without the need to posit any additional analogy-specific machinery. The results also parallel findings from the developmental literature on analogy, demonstrating a shift from an initial reliance on surface feature similarity to the use of relational similarity later in training. Variants of the model allow us to consider and rule out alternative accounts of its performance. We conclude by discussing how these findings can potentially refine our understanding of the processes that are required to perform analogical inference.

Keywords: analogy; inference; relational reasoning; development; connectionism; neural network

1. Introduction

In the past three decades, there has been a growing appreciation for the view that analogy lies at the core of human cognition (Gentner, 1983, 2010b; Hofstadter, 2001; Holyoak, Gentner, & Kokinov, 2001; Lakoff & Johnson, 1980; Leech, Mareschal, & Cooper, 2008; Penn, Holyoak, & Povinelli, 2008). Proponents of this view argue that it is our capacity to understand, produce, and reason with analogies (and metaphors) that allows us to create the wonderfully rich and sophisticated intellectual and cultural worlds we inhabit. For example, analogical reasoning can drive learning and cognitive development (Carey, 2009; Gentner, 2010a), facilitate abstract problem-solving and creative thinking (Gick & Holyoak, 1980; Holyoak & Thagard, 1996; Welling, 2007), and even play a foundational role in the scientific enterprise itself (Brown, 2003; Dunbar, 1995, 1997; Nersessian, 1992). For all these tasks, it is crucial to be able to draw an analogy between a source and target domain based on shared relational structure (e.g. solar systems and atoms both consist of a larger central object orbited by smaller satellite objects), rather than (or in the face of) shared surface features alone (e.g. tangerines and suns are both round and orange; Gentner, 1983).

In light of the varied, powerful ways in which analogy can contribute to human cognitive achievement, a great deal of theoretical, computational, and empirical work has been carried out in an attempt to understand the nature and structure of analogical reasoning and the cognitive

*Corresponding author. Email: pthibode@oberlin.edu

mechanisms that support it (for review, see French, 2002; Gentner & Forbus, 2011). The most common and most influential approach has been to decompose analogy into several sub-processes and to explore how these component processes operate together in analogical reasoning and inference. In a recent review, Gentner and Forbus (2011) describe four major sub-processes of analogy:[1]

(1) *Retrieval*: Given a situation, find an analog that is similar to it.
(2) *Mapping*: Given two situations, align them structurally to produce a set of correspondences that indicate 'what goes with what', candidate inferences that follow from the analogy, and a structural evaluation score which provides a numerical measure of how well the base and target align.
(3) *Abstraction*: The results of comparison may be stored as an abstraction, producing a schema or other rule-like structure.
(4) *Rerepresentation*: Given a partial match, people may alter one or both analogs to improve the match.

(Gentner & Forbus, 2011, p. 267)

This type of functional decomposition has inspired a great deal of behavioural research on analogical reasoning and has significantly increased our understanding of this complex cognitive process. For instance, researchers have been able to isolate some of the individual characteristics of each of these sub-processes in human analogical processing. *Retrieving* analogical matches from memory is largely driven by surface feature commonalities, while *mapping* is largely driven by relational commonalities (Forbus, Gentner, & Law, 1995). These insights have helped psychologists explain, among other things, when and why analogy may or may not be effectively deployed in problem-solving (Blanchette & Dunbar, 2000; Chen, 1996; Holyoak & Koh, 1987).

This divide-and-conquer approach has also furthered our understanding of analogical processing through its influence on computational models of analogy. The models that have been among the most popular and successful at simulating behavioural data in recent years – models like the Structure-Mapping Engine (SME; Falkenhainer, Forbus, & Gentner, 1989) and Learning and Inference with Schemas and Analogies (LISA; Hummel & Holyoak, 1997, 2003) – explicitly implement at least some of the key individual sub-processes outlined above as distinct systems that come together to give rise to analogical reasoning. In particular, these models of analogy share a basic commitment to treating questions of conceptual representation and mapping as separate issues. In both SME and LISA, this is achieved by constructing structured symbolic (or hybrid) conceptual representations (e.g. of objects and relations) and implementing a distinct mechanism dedicated to mapping (and/or binding) that operates over these representations (Gentner & Forbus, 2011; but see Leech et al., 2008).

The success that these models have had at simulating a wide range of behavioural findings suggests that they may capture certain important features of human analogical processing (Gentner & Forbus, 2011; Hummel & Holyoak, 2005). Indeed, this success has led some researchers to argue that the *algorithmic* processes which allow these models to exhibit analogical behaviour, namely structured representations and explicit mapping mechanisms, are *necessary* for any system to carry out analogical reasoning (Doumas, Hummel, & Sandhofer, 2008; Gentner & Markman, 1993; Hummel, 2010). Hummel (2010), for instance, argues that 'symbol systems permit qualitatively different kinds of processing (such as learning and inference) than do nonsymbolic systems (a difference so profound that our symbolic species dominates the planet, whereas our nonsymbolic cousins do not)' (p. 961). We refer to this perspective as the *Structural* approach to analogy.

The *Structural* approach constitutes an a priori constraint on any theory or model of analogy and implies that there are classes of models (and cognitive systems) that cannot *in principle* capture analogical reasoning (namely, any model that does not include structured representations

and a mapping mechanism). In particular, models that rely on low-level associative learning mechanisms and fully distributed representations to simulate cognitive processing, like many connectionist models (e.g. Rogers & McClelland, 2008), are thought to be incapable of supporting fully analogical capabilities (Holyoak & Hummel, 2008; Hummel, 2011; Kemp & Tenenbaum, 2008; Marcus, 2001; Marcus & Keil, 2008; Opfer & Doumas, 2008). A common view states that while non-symbolic models 'excel at learning complex correlations between features, they fail to represent abstract operations over variables, structured representations, and contrasts between individuals and kinds; and it is not clear how well they can do any of these things in principle' (Marcus & Keil, 2008, p. 722). Further, it has been argued that because these types of models do not represent relations and relation-filler bindings explicitly, they 'cannot use relational information to drive inference' (Opfer & Doumas, 2008, p. 723). Some theorists have gone so far as to suggest that the attempt to capture higher level cognitive abilities such as analogy in these sorts of models involves 'suck[ing] the essence out, then force-fit[ting] what's left into an associationist straitjacket' (Holyoak & Hummel, 2008, p. 389).

To summarise, then, the *Structural* approach embodies two related hypotheses: (1) that the subcomponents of analogy outlined above, especially a specific mapping mechanism that operates over structured representations, are necessary for analogical processing and (2) that models of a certain class (i.e. in which these components are not explicitly built in, or in which distributed representations and low-level learning mechanisms carry the burden of processing) cannot *in principle* come to instantiate these processes and therefore cannot implement analogical processing.

However, some models of the sort implicated in hypothesis (2) are known to exhibit complex emergent behaviour (Elman, 1990; Rogers & McClelland, 2008; St. John, 1992; Thelen & Smith, 1998). It is not always obvious what high-level functions such models can implement (or approximate) through the operation of low-level, general learning mechanisms over the course of development (Chalmers, French, & Hofstadter, 1992). Many researchers regard this as a key strength of these models, insofar as this allows for a more direct comparison and connection between what we know about the structure and function of the nervous system and the emergence of complex cognitive functions over the course of development (see, for instance, Elman et al., 1996; Spencer, Thomas, & McClelland, 2009). Therefore, we view it as an *empirical* rather than a *logical* question whether analogical processes can in fact emerge over the course of learning in one of these models. We refer to attempts to capture analogy in these sorts of models as the *Emergent* approach.

Critically, the *Emergent* framework denies that analogical reasoning is supported by symbolic conceptual representations and a distinct mapping or binding mechanism that operates over these symbols. In the *Emergent* framework, questions of mapping are inseparable from questions of conceptual representation. Therefore, the concept of *mapping* may be thought of as a computational-level description (Marr, 1982) of a fundamentally integrated phenomenon, but without direct implications for the algorithmic or implementation levels. On this view, the way that concepts are stored and represented in the system (i.e. that they are interrelated or even overlapping) may naturally give rise to analogical mapping over the course of development (French, 1995; Hofstadter, 1996).

Previously, proponents of the *Emergent* approach have highlighted the importance of studying the relationship between conceptual representation and relational reasoning. The Copycat and Tabletop models of analogy proposed by Hofstadter, Mitchell, and French (French, 1995; Hofstadter & Mitchell, 1994; Mitchell, 1993) describe relational reasoning as a type of generalisation that results from the blending of conceptual domains. Proponents of the *Emergent* approach have also had success simulating certain aspects of the development of analogical reasoning (see especially Leech et al., 2008).

Our work attempts to advance the *Emergent* framework by illustrating how a simple feedforward connectionist model can give rise to certain forms of analogical inference without recourse to specific mapping or binding mechanisms. Our approach, like the Copycat model, emphasises

the deep connection between questions of conceptual representation and relational inference. But unlike the Copycat model, our network learns conceptual representations through experience; in the Copycat model, conceptual representations and connections between representations were not learned. Further, unlike Copycat, which implemented conceptual representations as unitary nodes with excitatory and inhibitory connections to related conceptual nodes, our network represents conceptual knowledge in a truly sub-symbolic fashion, as distributed patterns of activation in intermediate layers of the network. Our work also builds on the insights of Leech et al. (2008) by focusing on knowledge-based analogical inference. A major criticism of Leech et al.'s (2008) work has been its failure to capture our ability to use analogy to drive inferences and facilitate knowledge acquisition (e.g. Markman & Laux, 2008; Opfer & Doumas, 2008).

In support of the *Emergent* approach, we use a computational framework that does not require classical structured representations or an explicit mapping mechanism. The framework is based on the Rumelhart model (Rumelhart, 1990; Rumelhart & Todd, 1993), used by Rogers and McClelland (2004, 2008), to investigate the development and representation of general semantic knowledge. In most of the simulations presented below, the network is given a partial phrase that it tries to complete. At the outset, the model does not 'know' anything about the symbols that constitute the phrase or how these symbols relate to one another. However, over the course of training, the model learns to represent these inputs and outputs, and their relationships to one another, in intermediate 'hidden' or 'representation' layers through an error-driven process of progressive differentiation.

Importantly, although the inputs and outputs are presented as individual or combinations of unitary nodes (symbols), the network is not symbolic in the traditional sense. The network does not perform operations over the input, output, or relational symbols. Instead, the semantic information is stored in the weights between the representation layers and, as we will see, relational reasoning emerges from the overlapping, distributed representations that are learned in the hidden layers.

This modelling approach is consistent with an account of learning in which children (and adults) are constantly making predictions about what they will experience in the world, and using their observations (i.e. what they actually do experience) to make better predictions in the future (Elman, 1991; Rogers & McClelland, 2004). This model has succeeded in capturing many results from the literature on semantic development in children (Rogers & McClelland, 2004, 2008), and we have previously applied it to simulate the development of conceptual metaphor, which is closely related to certain aspects of analogy (Flusberg, Thibodeau, Sternberg, & Glick, 2010; Thibodeau, McClelland, & Boroditsky, 2009; see also the response to commentators in Rogers & McClelland, 2008, for an earlier attempt to showcase some of the analogy-like abilities of this model).

The importance of the research presented here depends on two key considerations: (1) that the Rumelhart model is relevant to the discussion of the nature and structure of analogical reasoning and (2) that the tasks we ask the model to perform are indeed relevant, appropriate analogical reasoning tasks. Consideration (1) is relatively uncontroversial because the Rumelhart model has been explicitly (and implicitly) singled out as incapable of supporting analogical processing as a matter of a priori fact by proponents of the *Structural* approach, as detailed above. Consideration (2) may be a point of more contentious debate, an issue we turn to in greater depth in the general discussion.

We show that there are two key features of analogical reasoning that this model will come to exhibit over the course of learning. First, we demonstrate that the model can make an inference based on shared relational structure between a source and target domain in the face of conflicting feature-based similarity, a hallmark of mature analogical reasoning (Gentner, 1983). Second, we show that the model can leverage shared relational structure between two domains in order to facilitate learning (Carey, 2009; Gentner, 2010a). We also review other aspects of the model's behaviour that parallel data on the development of analogical reasoning, including a shift over time from making inferences based on shared surface features to inferences based on shared relational similarity (Gentner, 1988; Goswami, 1992; Rattermann & Gentner, 1998).

These findings leave proponents of the *Structural* approach to analogy with something of a dilemma. One possibility is to conclude that the model is succeeding at these analogical reasoning tasks by actually implementing processes like mapping, but doing so in an emergent, graded, and approximate fashion rather than an explicit and rule-like fashion (see, e.g., Frank, Haselager, & van Rooij, 2009; Monner & Reggia, 2012). This is our preferred interpretation, and we utilise the concept of *structured pattern completion* to help explain the behaviour of the network (Gentner & Markman, 1993, 1995). Indeed, the heart of this research project consists of exploring aspects of the model that allow it to implement these analogical processes, including architectural and learning-based constraints, as well as the important role language might play in the development of analogical reasoning (McClelland, 2009, 2010).

Another possibility, alluded to above, is to conclude that the model is not doing analogy at all. However, this option implies that some behaviours that have previously been claimed to require analogical reasoning (e.g. drawing certain inferences based on shared relational structure) do not actually require all the sub-components previously proposed and instead can be performed in some other way. We thoroughly discuss the implications of both of these possibilities in the general discussion. Either conclusion represents an advance in our understanding of the nature of analogical reasoning.

A brief outline of the remainder of the paper is as follows. We will first describe the general structure of the Rumelhart model architecture and introduce our task: learning about two families with alignable relational structure, each including parents, children, and a dog. After learning about the two families, we ask the model to answer a particular question about one of the families ('Who walks the dog?'). Critically, this test does not involve any fact that the model was explicitly trained on; the model can only resolve it by exploiting the relational similarity between the two families. A series of simulations allows us to explore aspects of its learning and rule out various alternative explanations for its performance. We then highlight the strengths and weaknesses of this modelling approach, including what features of analogical reasoning this type of model can simulate, and what behaviours might require additional cognitive machinery. We conclude with a discussion of how the *Emergent* approach to analogy relates to existing theories and address some of the long-standing critiques of the (in)ability of sub-symbolic models to learn relational structure, along with the other issues raised above.

2. General modelling framework

Our learning task is inspired by Hinton's (1986) family tree model, one of the first attempts to address relational learning in a connectionist network. Previous empirical work has shown that family trees are closely related to analogy, to the point that practice with family trees facilitates analogical inference in young children (e.g. Mutafchieva & Kokinov, 2007). The goal of the model is to learn 'statements' that are true about the various members of a family, including identity information, perceptual features, and relations between family members. The input to the model consists of activating a Subject unit, corresponding to a particular family member, and a Relation unit. The Relation units correspond to the different kinds of relationships that can hold between subjects and objects (e.g. 'is_named', 'is_a', 'has', 'parent_of', and 'daughter_of'). The network is wired up in a strictly feed-forward fashion, as shown in Figure 1, such that the input propagates forward through the internal layers, resulting in a set of predictions over the Object layer. Over the course of training, the network's weights change (via backpropagation of the cross-entropy error on the output units) in order to better predict which Objects hold for each particular Relation to each Subject. As the model also contains intervening layers of units between the input and output layers, it is forced to re-represent the inputs as a distributed pattern of activation over these internal layers.[2]

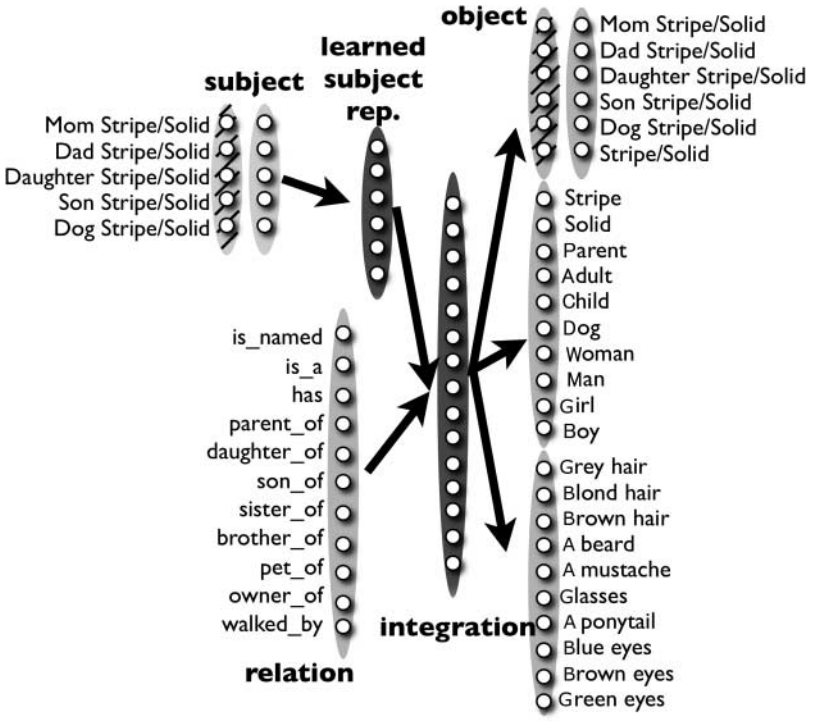

Figure 1. The network architecture. Note that in some cases more or fewer units were used in the Subject, Relation, or Object layer to accommodate more or fewer families or family members.

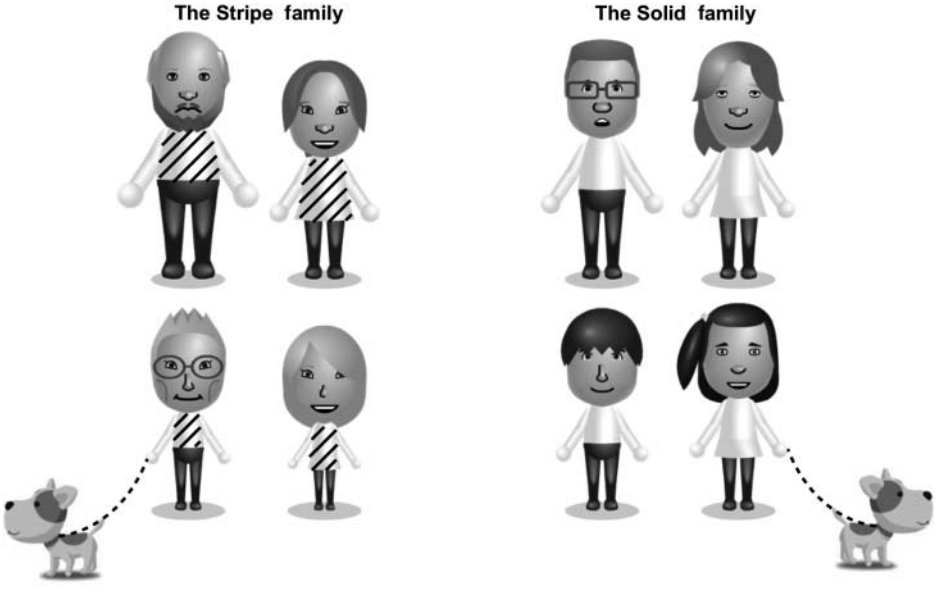

Figure 2. An illustration of the Stripe and Solid families, which served as the source and target domain for most simulations.

In this simple world, the network learns about two families: the Stripes and the Solids (pictured in Figure 2 and described in Table 1). Each family has a mother, a father, a daughter, a son, and a pet dog. Each individual has an identity (e.g. 'Daughter' and 'Stripe') given by the 'is_named' relation.

In these training patterns there is no output overlap between the families. Specifically, there is one unit that represents membership in the Stripe family and one that represents membership in the Solid family. In addition, there is one output unit associated with each individual that represents their name.

Each individual can be described as belonging to a variety of categories – for example, human, dog, parent, child – given by the 'is_a' relation. In this context, corresponding family members

Table 1. Example set of training patterns with agent/subject (column 1), relation (column 2), and completion/object (column 3).

$\text{Mom}_{\text{Stripe}}$	*is_named*	Stripe, $\text{Mom}_{\text{Stripe}}$
$\text{Dad}_{\text{Stripe}}$	*is_named*	Stripe, $\text{Dad}_{\text{Stripe}}$
$\text{Daughter}_{\text{Stripe}}$	*is_named*	Stripe, $\text{Daughter}_{\text{Stripe}}$
$\text{Son}_{\text{Stripe}}$	*is_named*	Stripe, $\text{Son}_{\text{Stripe}}$
$\text{Dog}_{\text{Stripe}}$	*is_named*	Stripe, $\text{Sog}_{\text{Stripe}}$
$\text{Mom}_{\text{Stripe}}$	*is_a*	Stripe, human, adult, parent, mom
$\text{Dad}_{\text{Stripe}}$	*is_a*	Stripe, human, adult, parent, dad
$\text{Daughter}_{\text{Stripe}}$	*is_a*	Stripe, human, child, daughter
$\text{Son}_{\text{Stripe}}$	*is_a*	Stripe, human, child, daughter
$\text{Dog}_{\text{Stripe}}$	*is_a*	Stripe, dog
$\text{Mom}_{\text{Stripe}}$	*has*	Grey hair, ponytail, brown eyes
$\text{Dad}_{\text{Stripe}}$	*has*	a bald head, beard, brown eyes
$\text{Daughter}_{\text{Stripe}}$	*has*	blond hair, blue eyes
$\text{Son}_{\text{Stripe}}$	*has*	blond hair, glasses, brown eyes
$\text{Dog}_{\text{Stripe}}$	*has*	blond hair, fur, brown eyes
$\text{Mom}_{\text{Stripe}}$	*parent_of*	$\text{Son}_{\text{Stripe}}$, $\text{Daughter}_{\text{Stripe}}$
$\text{Dad}_{\text{Stripe}}$	*parent_of*	$\text{Son}_{\text{Stripe}}$, $\text{Daughter}_{\text{Stripe}}$
$\text{Daughter}_{\text{Stripe}}$	*daughter_of*	$\text{Mom}_{\text{Stripe}}$, $\text{Dad}_{\text{Stripe}}$
$\text{Son}_{\text{Stripe}}$	*son_of*	$\text{Mom}_{\text{Stripe}}$, $\text{Dad}_{\text{Stripe}}$
$\text{Daughter}_{\text{Stripe}}$	*sister_of*	$\text{Son}_{\text{Stripe}}$
$\text{Son}_{\text{Stripe}}$	*brother_of*	$\text{Daughter}_{\text{Stripe}}$
$\text{Dog}_{\text{Stripe}}$	*pet_of*	$\text{Son}_{\text{Stripe}}$
$\text{Son}_{\text{Stripe}}$	*owner_of*	$\text{Dog}_{\text{Stripe}}$
$\text{Dog}_{\text{Stripe}}$	*walked_by*	$\text{Son}_{\text{Stripe}}$
$\text{Mom}_{\text{Solid}}$	*is_named*	Sollid, $\text{Mom}_{\text{Solid}}$
$\text{Dad}_{\text{Solid}}$	*is_named*	Solid, $\text{Dad}_{\text{Solid}}$
$\text{Daughter}_{\text{Solid}}$	*is_named*	Solid, $\text{Daughter}_{\text{Solid}}$
$\text{Son}_{\text{Solid}}$	*is_named*	Solid, $\text{Son}_{\text{Solid}}$
$\text{Dog}_{\text{Solid}}$	*is_named*	Solid, $\text{Dog}_{\text{Solid}}$
$\text{Mom}_{\text{Solid}}$	*is_a*	Solid, human, adult, parent, mom
$\text{Dad}_{\text{Solid}}$	*is_a*	Solid, human, adult, parent, dad
$\text{Daughter}_{\text{Solid}}$	*is_a*	Solid, human, child, daughter
$\text{Son}_{\text{Solid}}$	*is_a*	Solid, human, child, daughter
$\text{Dog}_{\text{Solid}}$	*is_a*	Solid, dog
$\text{Mom}_{\text{Solid}}$	*has*	Grey hair, green eyes
$\text{Dad}_{\text{Solid}}$	*has*	Grey hair, moustache, glasses, brown eyes
$\text{Daughter}_{\text{Solid}}$	*has*	Brown hair, ponytail, green eyes
$\text{Son}_{\text{Solid}}$	*has*	Brown hair, green eyes
$\text{Dog}_{\text{Solid}}$	*has*	Brown hair, fur, brown eyes
$\text{Mom}_{\text{Solid}}$	*parent_of*	$\text{Son}_{\text{Solid}}$, $\text{Daughter}_{\text{Solid}}$
$\text{Dad}_{\text{Solid}}$	*parent_of*	$\text{Son}_{\text{Solid}}$, $\text{Daughter}_{\text{Solid}}$
$\text{Daughter}_{\text{Solid}}$	*daughter_of*	$\text{Mom}_{\text{Solid}}$, $\text{Dad}_{\text{Solid}}$
$\text{Son}_{\text{Solid}}$	*son_of*	$\text{Mom}_{\text{Solid}}$, $\text{Dad}_{\text{Solid}}$
$\text{Daughter}_{\text{Solid}}$	*sister_of*	$\text{Son}_{\text{Solid}}$
$\text{Son}_{\text{Solid}}$	*brother_of*	$\text{Daughter}_{\text{Solid}}$
$\text{Dog}_{\text{Solid}}$	*pet_of*	$\text{Daughter}_{\text{Solid}}$
$\text{Daughter}_{\text{Solid}}$	*owner_of*	$\text{Dog}_{\text{Solid}}$
$\text{Dog}_{\text{Solid}}$	*walked_by*	??? ($\text{Daughter}_{\text{Solid}}$)

Note: The shaded pattern is the test pattern: it is omitted in training and presented in test to see if the network can correctly infer who walks the Solids' dog.

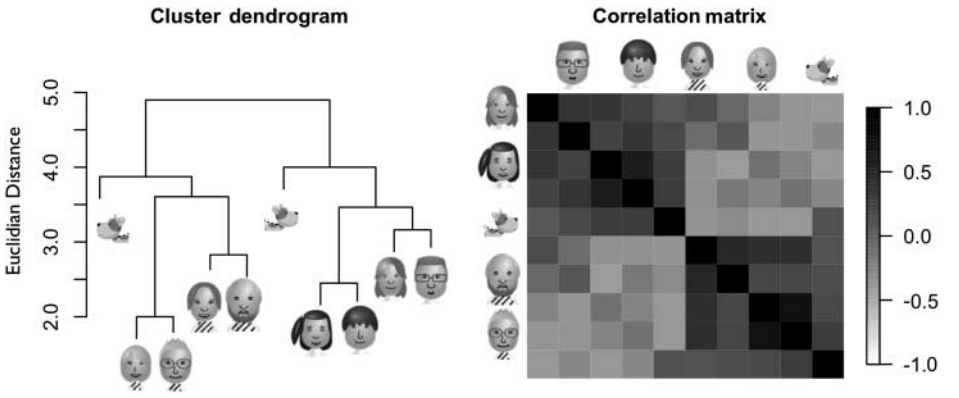

Figure 3. The left panel displays a hierarchical cluster of the training patterns used for Simulation 3. The right panel displays a correlation matrix of these same training patterns. The training patterns are structured such that any given member of either family is most similar to another member of the same family. Additionally, despite mostly non-overlapping inputs and outputs, the families are structured analogously.

in the two families share numerous features. For instance, the mother in the Stripe family and the mother in the Solid family are both 'human', 'parent', 'woman', etc.

Each individual has a set of perceptual features, such as grey hair, a moustache, or glasses, given by the 'has' relation. In this context, there is also feature overlap across families as well as feature overlap within families. For instance, the mother in the Stripe family and the mother in the Solid family both have grey hair, while the daughter in the Stripe family and the son in the Stripe family both have blonde hair.

Finally, the members of each family sit in various relations to one another. For example, the mother in the Stripe family is the 'mother_of' the son and daughter in the Stripe family. While there is between-family overlap, across all training patterns in all contexts the similarity of each person to all their family members is greater than to any non-family member (Figure 3).

Of note, this implementation of relations differentiates our *Emergent* approach from *Structured* models of analogy like SME and LISA. In Structured models, relational and featural information are treated as fundamentally different kinds of information: relations are operators over internal, feature-based representations (Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 1997, 2003). In contrast, in our approach, the model's learned representations of the input items integrate these two types of information, treating relations as a fundamental part of the structure of experience.

Both of these ways of thinking about and implementing relations may be important for analogical reasoning, although they may underlie our thinking about different kinds of analogies. Some analogies may require explicit mapping across domains – explicitly identifying relational and featural correspondences in different domains (Holyoak, Novick, & Melz, 1994) – whereas others may be more implicit and automatic. Implementing relations as part of experience may be consistent with this latter class of analogies, which have received relatively less attention in the analogy-modelling literature. We elaborate on this proposed distinction between different kinds of analogies in the general discussion (but see Rogers & McClelland, 2004, 2008, for a more detailed account of the advantages of treating relations as part of experience).

The underlying model parameters were identical in all the simulations that we present unless otherwise indicated. The learning rate was 0.005 and the network was trained for 10,000 epochs. Results were averaged over 10 runs of each network in order to provide data for statistical tests and rule out the possibility that our results are driven by a random, unlikely configuration of network weights. The hidden layers consisted of 6 Subject Representation units and 16 Integration units.

In all presented simulations, error on the training patterns was very low by the end of training (average cross-entropy error <0.35).

3. Simulations

3.1. *The basic model*

In the first simulation, the network learns about the Stripes and the Solids, with a single fact omitted about the Solid family. However, the families are otherwise very similar, with isomorphic relational structures (i.e. they have the same family members in the same relationships to one another). In particular, the daughter of the Stripes both owns the dog and walks the dog. While the network knows that the daughter of the Solids owns their dog, it receives no information about who walks their dog. This network does a good job of learning the facts on which it is trained, but the question of interest is whether it can extend its knowledge to answer a question on which it received no training: Who walks the Solids' dog?

We can contrast two major predictions. Naively, one might think that the network runs on raw association. As the Solids' dog is most similar to the Stripes' dog, the network should conclude that the Stripes' daughter walks the Solids' dog! Alternatively, we might expect that the network will encode the relational structure between the two families, and so will correctly conclude that the person in the appropriate position within the Solid family – namely, the Solid's daughter – will be the one who walks their dog. In fact, the network decides that within the Solid family, the daughter walks the dog. A paired t-test contrasting the activation levels of the Stripes' daughter with the Solids' daughter is highly significant, $t[9] = 7.75$, $p < .001$ (Figure 4).

It might be the case that the network is not driven by the relational similarity between the two families, but rather by some non-obvious feature of the input within the Solid family alone. For example, perhaps the fact that the daughter owns the dog creates enough of an association that the network would also conclude that she walks the dog, even without drawing any analogical inferences from the Stripe family. To control for this, we ran a second simulation, in which the model was trained only on the Solid family, with no information about the Stripe family. In this network, the model does not conclude that the daughter walks the dog. Instead, it defaults to a different kind of mapping on which it has also been trained: the identity mapping. Without any

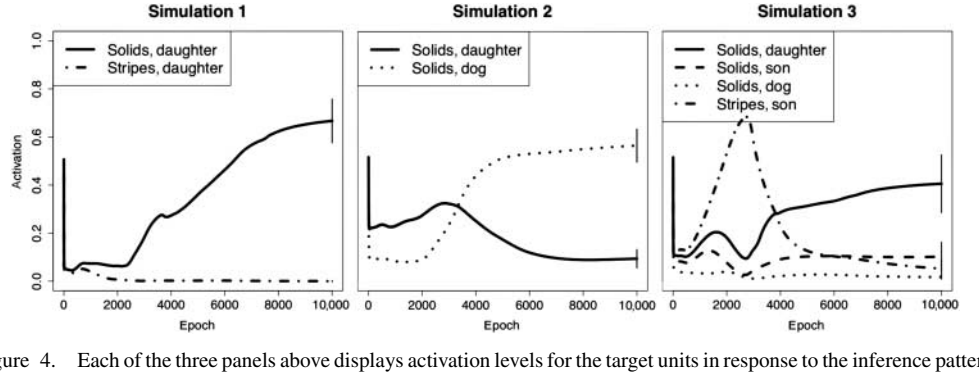

Figure 4. Each of the three panels above displays activation levels for the target units in response to the inference pattern over the course of learning. The left panel illustrates results from the first simulation when the daughter walks the dog in both the Stripe and Solid families; the middle panel illustrates results from the second simulation when the model only learns about the Solid family; and the right panel illustrates results from the third simulation when the son walks the dog in the Stripe family and the daughter walks the dog in the Solid family.

further information, the network decides that the dog walks itself! A paired t-test shows that activation of the Solids' dog is greater than the Solids' daughter, $t[9] = 5.61, p < .001$ (Figure 4).

Simulations 1 and 2 do not, however, distinguish another set of predictions. One possibility is that the network has learned to align the two families with respect to their relational structure, but only in an exact way. On this account, the model may have placed both mothers, both daughters, and both dogs in exactly the same structure, perhaps a tree structure, with a dimension dividing the families from each other, driven by the *overall* relational similarity between the families. Another possibility is that the network learns to associate certain common features, such as the features shared among the analogous members of each family, in order to drive its success on the relational questions. On either of these views, the network should only be able to align the structures between the two families when the correspondence is exact, or nearly so.

On the other hand, perhaps the network has learned the details of the family relations within each family as well as across families. In this case, it could learn a regularity like 'whoever owns the dog, walks the dog', which is driven neither by perfect, global structural alignment nor by associations between surface features. This kind of relational generalisation is closely related to those tasks that previous researchers have argued can only be done using a distinct mapping mechanism operating over explicit symbols (Gentner & Markman, 1993, 1995; Holyoak & Hummel, 2000; Hummel, 2010; Markman, 1999). Therefore, it would be a surprising and exciting finding if this network were able to succeed in such an abstract relational mapping task.

In order to distinguish between these hypotheses, we ran a third simulation, very similar to the first, except that in the Stripe family, the *son*, not the daughter, both owns and walks the dog. When the network is informed that the daughter of the Solids owns the dog, but is not told who walks it, what should the network conclude? If the network needs to align each member of each family exactly, it should overlay the two dogs in the same place relative to each family and conclude that the Solids' son walks their dog. Similarly, if overall association of the dog to certain features (perhaps those that the sons share) is driving learning, then the Solids' son should again walk the dog. However, if the network is learning the details of the relational structure, and in particular the regularity between owning a dog and walking it, then the network should succeed in inferring that the Solids' daughter walks the dog.

This is precisely what occurs. Separate tests contrasting the activation level of the Solids' daughter with the activation level of the Stripes' son, $t[9] = 2.58, p < .05$, the activation level of the Solids' son, $t[9] = 2.95, p < .05$, and the dog, $t[18] = 3.35, p < .01$, are all significant (Figure 4). This demonstrates that raw co-occurrence or other simple associative processes which are often believed to underlie the performance of error-driven learning models (e.g. Hummel, 2010, in reply to Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010) are not the key to learning in this model. It is, however, interesting to notice that the Stripes' son is the model's choice early in training, suggesting that the network first tends to make judgements predominately based on surface similarity, but over time shifts towards judgements based on relational similarity. This 'relational shift' has been widely observed in the literature on the development of analogical reasoning abilities (Gentner, 1988; Goswami, 1992; Rattermann & Gentner, 1998). Intriguingly, this pattern is observed throughout the various simulations presented in this paper.

3.2. *Extending the model*

We have presented a basic set of simulations showing that the model succeeds in performing analogical inference from a family that is fully described (and learned) to one that is less fully described. In the simulations below, we will extend the basic model in several directions, addressing possible objections to our claim that the model really is succeeding at analogical inference. Each of these models will extend the third simulation, in which the son of one family owns and

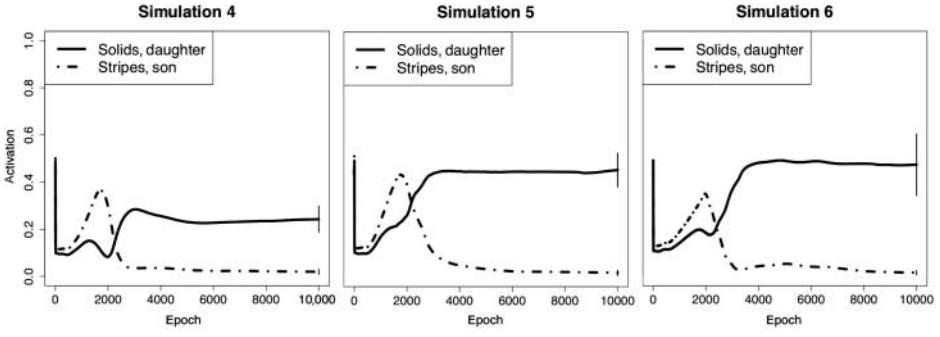

Figure 5. The three panels above display activation levels for the target units for Simulations 4–6 over the course of learning. The left panel illustrates results from the fourth simulation in which the family trees have relational structures that are less readily alignable; the middle panel illustrates results from the fifth simulation in which the network takes distributed input representations of the family members; and the right panel illustrates results from the sixth simulation in which there are no overlapping output units.

walks the dog, and the task of the model is to infer that the daughter of the other family, who owns the dog, also walks it.

3.2.1. *Inexact match: Can the model align non-isomorphic structures?*

In the previous simulations, each family had the same number of family members, sitting in the same (or extremely similar) relationships to one another. We can investigate the extent to which the network relies on perfectly overlaying the two families by making the family structures only approximately match. In the fourth simulation, the Stripes have three children, two sons and a daughter, and one of the sons again owns (and walks) their dog. The Solid family still has two children, one son and one daughter, and their daughter owns the dog. The model continues to make the inference that she probably walks the dog as well. A paired t-test contrasting the activation levels of the Solids' daughter with the Stripes' son is highly significant, $t[9] = 4.28$, $p < .01$ (Figure 5). This demonstrates that the network can learn to draw inferences over structures that are only partially alignable, which has been shown to be important for analogical reasoning in previous work (such as Falkenhainer, Forbus, & Gentner, 1989).

3.2.2. *Distributed inputs: Does the model rely on implementing symbols?*

We have claimed that the success of this network depends on its development of distributed, sub-symbolic representations, with which it can integrate the perceptual and the relational information about the family members within a high-dimensional representational space. Others might argue instead that the network is simply implementing symbols and succeeds by performing some syntax-like transformation on those symbols. Such an argument may point to the localist input units that represent the family members. We argue that the localist inputs are a useful simplification, but that focusing on them is a distraction, as the network can never directly exploit these localist units. Instead, it is required to re-represent each item as a pattern of activation over a hidden layer, as described above.

To make this point more clear, we ran a fifth simulation, which included distributed, rather than localist, input representations for the family members. Following a model by Rogers and McClelland (2004), these were simply chosen to be the perceptual features of each family member. For instance, whereas in the first four simulations the mother in the Stripes family was represented by a localist unit corresponding to her identity, in the fifth simulation she was represented by

a collection of units (e.g. 'grey hair', 'ponytail', and 'brown eyes') that describe her physical appearance. This should not assist the network in acquiring the relational structure; if anything, it might appear to bias the network towards using surface-level perceptual features for generalisation. Nevertheless, the network still infers that the owner of the dog walks it, transferring from the Stripes' son to the Solids' daughter. A paired t-test contrasting the activation levels of the Solids' daughter with the Stripes' son was statistically significant, $t[9] = 6.05$, $p < .001$ (Figure 5).

3.2.3. *Non-overlapping inputs: Does the model require perceptual overlap?*

On the other hand, one might argue that the architecture is biased in the opposite direction: the more direct overlap between the two families at the feature level (i.e. the output layer), the less work the model needs to do to align their structures. What if only the relational similarity is available, as might be the case when constructing analogical mappings across very different domains of knowledge? This kind of analogy may be critical for explaining how analogy can subserve cognition and reasoning more generally (e.g. Gentner, 2010b).

To test this, we carried out a sixth simulation in which the training patterns for the two families had completely non-overlapping output units. The network essentially had two copies of each output property, so that each family's target representations were totally distinct. To succeed in generalising the relation between the two families, the network would need to align the structures even in the absence of *any* surface-level similarity between the two families. And this is precisely what it did. Again, when the network is told that, in the Stripe family, the son owns and walks the dog, it concludes that for the Solids, the owner of the dog – the daughter – must also walk it, as evidenced by a paired t-test contrasting the activation levels of the Solids' daughter with the Stripes' son, $t[9] = 3.58$, $p < .01$ (Figure 5).

3.2.4. *Scaling-up: Can the model make inferences when given more than two families?*

It remains to be shown that the ability of the model to make analogies does not depend on it living in a world with only two different structures. Is it able to extend its learning to multiple families? To investigate this question, we ran two different simulations, similar to those above, with four rather than two distinct families (adding the Dash family and the Dot family). In Simulation 7, a different member of each family owns and walks their dog: the father of one family, the mother of another, the son of the third, and, finally, the daughter of the target family. In Simulation 8, in addition to this variability, two of the families have slightly different structures: one has only a son, while the other has two sons and a daughter.

In both cases, the network infers that in the target family, the daughter must also walk the dog. For Simulation 7, a within-subjects ANOVA using a planned contrast comparing the activation values of the Solids' daughter with the Stripes' son, the Dashes' mother, and the Dots' father (each dog walker in their respective family) was significant, $F[1, 36] = 31.10$, $p < .001$. Paired t-tests contrasting the activation levels of Solids' daughter with the dog walkers in each of the other families, including the Solids' son $t[9] = 3.47$, $p < .01$, the Dashes' mother, $t[9] = 3.45$, $p < .01$, and the Dots' father, $t[9] = 3.47$, $p < .01$, were also significant (Figure 6). For Simulation 8, a within-subjects ANOVA using a planned contrast comparing the same activation values in the families with greater variability was also significant, $F[1, 36] = 42.40$, $p < .001$. Paired t-tests contrasting the activation levels of the Solids' daughter with the dog walkers in each of the other families, including the Solids' son, $t[9] = 7.39$, $p < .001$, the Dashes' mother, $t[9] = 7.37$, $p < .01$, and the Dots' father, $t[9] = 7.31$, $p < .001$, were significant (Figure 6).

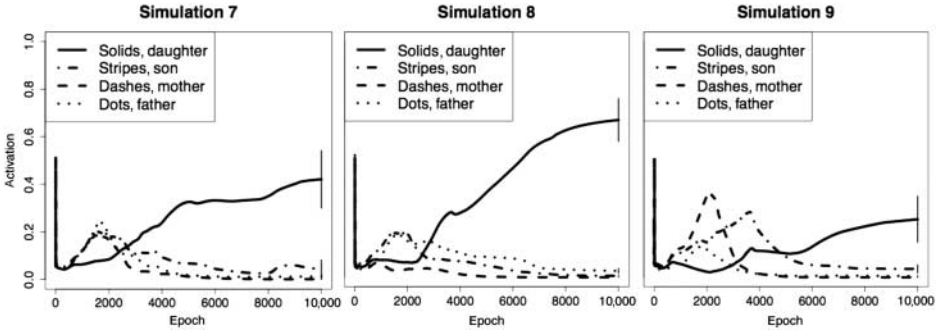

Figure 6. The left panel illustrates results from the seventh simulation in which there are four families, each with a different family member walking the dog. The middle panel illustrates results from the eighth simulation in which there are four families with relational structures that are less readily alignable. The right panel displays activation levels for the target unit in the ninth simulation in which the 'pet_of' relation was removed and additional Objects were 'owned'.

3.2.5. *Does the network merely have the specific relations and features required to solve the problems on which it is tested?*

In each of the simulations that we have presented so far, a suite of relations and features has facilitated the analogical mapping. In every case, there is only one 'owner_of' relation that uniquely links the dog in a given family to their owner (e.g. the Stripes' son and the Stripes' dog) and only one 'pet_of' relation that similarly links the dog in a given family to their owner (e.g. the Stripes' son and the Stripes' dog). It may be argued that the 'walked_by' relation simply mirrors the behaviour of these two relations.

We are not entirely opposed to this interpretation of the model's performance. Since the network never experiences two relations at the same time, it can never learn to associate relations through direct co-occurrence. Such mirroring, therefore, could be a result of extracting higher order contextual co-occurrence. As we pointed out in the introduction, this would be an interesting finding that many have argued is beyond the capability of sub-symbolic learning models. However, in order to rigorously test this possibility, the relational correspondences between the pet owners and their pets would need to be more complex and nuanced. That is, if the model can only learn to extract the relational correspondence between owning and walking a pet in situations where there are multiple relations to support the mapping and none to potentially misdirect the mapping, then the applicability of our approach would be extremely limited.

We ran a ninth simulation to explore more complex relational correspondences. In this simulation, we eliminated the 'pet_of' relation so the network was never given a family's dog and 'pet_of' as input. We also added several objects (a house, a car, toys, a doll, and a robot) that could be 'owned' along with the dog by various family members: moms and dads of every family owned 'houses' and 'cars', kids and dogs in every family owned 'toys', daughters owned 'dolls', and sons owned 'robots'. These training patterns were added and subtracted from those that were used in the eighth simulation, in which there were four families, each with slightly different kinship structures, and in which a different member owned and walked the dog in each family.

To test whether the network would still infer that the Solids' daughter walked the Solids' dog, we ran a within-subjects ANOVA using a planned contrast comparing the activation values of each dog walker in their respective family. The results of the ANOVA were significant, $F[1, 36] = 16.66$, $p < .001$, as were paired t-tests contrasting the activation levels of the Solids' daughter with the dog walkers in each of the other families, including the Solids' son, $t[9] = 2.50$, $p < .05$, the Dashes' mother, $t[9] = 2.20$, $p = .05$, and the Dots' father $t[9] = 2.50$, $p < .001$ (Figure 6). This suggests that the network can extract relational correspondences in more complex and varied contexts.

3.2.6. *Learning with structure: Can shared relational structure drive learning in this model?*

In the simulations presented above, we have explored some of the conditions under which a connectionist network can successfully perform analogical inference. By extending the learning environment in a variety of ways, we have also tried to anticipate and address objections to the specific claim that the network is performing analogical inference. Here, we explore whether the model can use analogy to facilitate the learning process itself. Researchers have demonstrated that people can leverage shared relational structure between domains to drive learning (e.g. Carey, 2009), and others have suggested that analogy plays a powerful, if not critical, role in cognitive development more generally (Gentner, 2010a). Can our model learn about a new domain of knowledge *faster* by relying on shared relational structure with a domain it already knows about?

Interestingly, this question also allows us to address a general critique that is commonly levelled against connectionist models: that they take too long to learn (e.g. Hummel & Holyoak, 2003). It has been argued that connectionist models are a poor representation of the human mind because they take, in the extreme case, thousands of exposures to learn what a human can learn on a single trial (Marcus, 2001). Indeed, in the simulations that we have already presented in this paper, the models were trained for 10,000 epochs to learn simple facts like 'the Stripes' son is the son of the Stripes' mother and father'. In contrast, it is argued, no person needs to be told that a son is the child of the son's mother more than a few times before committing this fact to memory. While this is certainly a relevant concern, we wish to point out a few reasons why this critique is, in our view, misguided.

First, it is important to spell out our commitment to the complementary learning systems approach – namely, the existence of distinct slow- and fast-learning systems in the human brain (McClelland, McNaughton, & O'Reilly, 1995).[3] There is ample behavioural, neurological, and computational evidence that the brain comprised a slow-learning cortical system and a fast-learning hippocampal system that function together in a complementary fashion to support memory and cognition (e.g., Eichenbaum, 2000; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). Following this framework, it is best to view our model as implementing a slow-learning system (McClelland & Rogers, 2003). Therefore, we do not view the fact that the network takes hundreds (or thousands) of epochs of training to learn as a drawback in itself. We believe there may be qualitatively different kinds of analogical inference and this slow-learning model may not be as useful for thinking about some cases of analogical reasoning (such as analogies that involve explicitly mapping the target and base domains, which may require working memory and cognitive control abilities). However, we feel that other cases of analogical inference are likely to be supported by this slow, developmental system, and understanding how this process emerges over developmental time is one of the goals of this paper. We return to this issue in the general discussion.

Second, we do not equate an epoch of training in a model with an exposure to an experience in real life. Instead, a single real-world experience likely results in multiple 'passes' through hippocampal and cortical networks in the brain. The recurrent nature of biological networks leads them to repeatedly cycle activation patterns (Cohen & Eichenbaum, 1993; McClelland & Goddard, 1996). On this view, even when a child exhibits something like 'fast-mapping', seemingly acquiring new knowledge in one trial, it is unlikely that her brain does a single feed-forward pass of activation and then re-wires itself to have encoded the new label. Rather, the new experience may be cycled and re-cycled through her brain during the dozens of seconds between exposure and test – with a learning algorithm getting hundreds of opportunities to impose some synaptic adjustments.

Third, while we train the model for 10,000 epochs in each simulation, it does not take the model 10,000 epochs to begin making the correct inference. As Figures 4–7 illustrate, the model reliably activates the correct output unit on the inference trial by about 2000–3000 epochs of training in most cases.

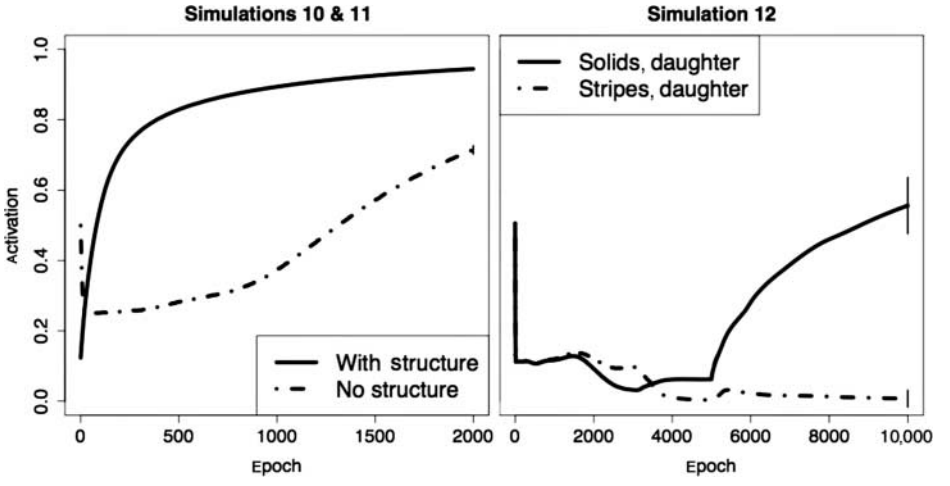

Figure 7. The left panel illustrates results from the 10th and 11th simulations in which knowledge was incorporated into existing structure or into a network without existing knowledge. The right panel displays activation levels for the target unit in the 12th simulation in which the model generalised to the target domain once it learned who walked the dog in the source domain.

Fourth, it is important to note that the model has absolutely no prior experience with anything at the outset of each simulation! It starts with a completely blank (and random) slate, other than the basic architectural constraints it embodies. People, on the other hand, almost always have relevant prior experience to build upon. One reason that people can easily associate a child with a parent is that they have a stable conception of what it means to be a child and what it means to be a parent. That is, learning to link a child to their parent involves integrating this new piece of information into an existing knowledge structure. This also seems likely to be the case for a neural network model. Here, we test this intuition about the network by exploring whether integrating novel information into an existing knowledge structure speeds learning in a connectionist model.

In the tenth simulation, we ran a modified version of the second simulation, in which the model was trained on the Solid family alone, with no information about the Stripe family and no information about who walks the dog. In this case, the model learns about the various family members and their relationships to one another until this aspect of the training environment is stably represented in the network (i.e. for 5000 epochs of training). Then, with this structure in place, the network is introduced to a novel fact about the family – that the Solids' daughter walks the dog.

Recall that when the network is trained on a single family and is not told who walks the dog, it thinks that the dog walks itself. After 5000 epochs of training, given the 'dog' and 'walked_by' input units, the network activates the 'dog' output unit most strongly (mean activation = .515, sd = .15) and the daughter unit only weakly (mean = .123, sd = .093).

On the 5001th epoch of training, we introduced the fact about the daughter walking the dog and then recorded how many epochs it took for the network to correctly activate the 'daughter' unit to a threshold of .5 when presented with the 'dog' and 'walked_by' input units. We found that, on average, it took 78 epochs (sd = 40.7) for the network to reliably learn that the Solids' daughter walks the dog when it had prior experience to build upon. At this point, activation of the 'dog' output unit had fallen considerably (mean activation = .240, sd = .087).

To contrast the speed at which the network learns who walks the dog when there is pre-existing structure about the family, we also ran an 11th simulation in which the target pattern was included as a training pattern from the beginning. In this 11th simulation, it took, on average, 1318.1 epochs (sd = 86.0) to reliably learn that the Solids' daughter walks the dog, which is significantly longer,

$t[12.8] = 125.0$, $p < .001$. The learning trajectory of this fact in the two simulations is plotted in Figure 7. When the model is integrating the fact about who walks the dog into pre-existing structure, it learns the fact more than 16 times faster than it does when starting from a blank slate.

Finally, we ran a 12th simulation to explore generalisation in this context. Here, we trained the network on two families – the Stripes and the Solids – for 5000 epochs on all the training patterns except those that specified who walks the dog in each family. On the 5001th epoch of training, we introduced a single new pattern: that the Stripes' daughter walks the Stripes' dog. We were interested in whether the model would generalise this information to its representation of the Solids' family. That is, would the shared relational structure between the domains cause the network to change its mind about who walks the Solids' dog in light of the information about who walks the Stripes' dog?

We found that it did! Activation of the Solids' dog output unit was significantly higher after 10,000 epochs of training than it was after 5000 epochs of training, $t[9] = 6.799$, $p < .0001$ (Figure 7). These findings suggest that the model can leverage shared relational structure to speed up learning and drive analogical inference.

3.3. *Discussion*

To summarise the results of the above simulations, we have demonstrated that analogical inference can emerge from a domain-general, distributed connectionist model of semantic learning and reasoning. Critically, this analogical inference (1) is driven by generalisation from a source domain to a target domain; (2) relies on abstract relational structure, not surface-level similarities or direct featural associations or co-occurrences; (3) parallels important features of the development of analogy in children; (4) can operate over structures which only approximately match, or which are only partially alignable; (5) exploits structural similarity even in the absence of explicit feature overlap, allowing the possibility of cross-domain analogical inference in guiding learning; (6) scales up to more complex training sets; and (7) can be learned (and is generalisable) relatively quickly, suggesting that the model can leverage shared relational structure to facilitate learning.

How is it that a connectionist model that lacks symbolic representations and an explicit mapping mechanism can succeed at this kind of analogical inference task? As we have demonstrated in several variations of the model, it is not due to any direct co-occurrence of features. Neither is it due to any kind of surface-level similarity between the items. In the extreme case, the two families can share absolutely no output units, and the model will still draw on the appropriate relational structure to make novel inferences. We argue that part of the answer involves the progressive differentiation of its representations over the course of development. Initially, all the weights are set to very small random values, so the network essentially treats every family member, and every relation, as being the same. Over the course of training, the model learns to 'pull apart' those representations that must be differentiated in order to produce the right answers. However, it only does so in response to erroneous predictions. One crucial constraint on this process that our model embodies is a particular architectural design that forces inputs from each domain to pass through the same sets of weights and hidden units. Any changes to the weights that influence one representation will also tend to affect similar representations. This biases the network to reuse as much representational structure as it can get away with. Alternative architectures that do not enforce the same general constraints fail to capture these patterns of learning (Rogers & McClelland, 2004).

In this particular network, the families share a great deal of structural similarity. If the family members are represented as points in a high-dimensional space, which is one approximate characterisation of the network's hidden representations, then the members of each family sit in similar positions relative to the other family (or families). That is, the father and mother in the Stripes bear the same relationship to each other and to their children as do the father and mother in the

Solids. An efficient representation of these similarities is to use one dimension to separate the families from each other, and the remaining dimensions to capture the relational structures common to each family (indeed, another model that learns family trees settles on exactly this kind of representation; Hinton, 1986).

This also applies to the network's representation of the relations. At the outset, the network does not 'know' that the relations 'sister of' and 'brother of' are more closely related than 'sister of' and 'has'. The network comes to learn the similarity structure of the relations through experience, aligning relations that are used in similar contexts (i.e. with similar Subjects and Objects).

As a result, the network's representations of the families become aligned over the course of training, because this allows the network to learn more efficiently (i.e. to reduce error more quickly). The side effect of this representational overlap is that when the network learns a fact about one family (e.g. one dog's owner walks it), the representations of the members of the *other* family (e.g. between that dog and its owner) get to come along for the ride. This is not to say that the model is stuck with its first guess about the structure of the world. As we indicated in the description of Simulation 3, and as is visible in other simulations as well, the model undergoes a developmental shift from predominantly perceptual to predominantly relational inference when the environment warrants such a shift (a finding we discuss in more detail below).

We can observe the process of progressive differentiation in this network by looking at a clustering diagram and a correlation matrix across the Subject Representation layer at different points in time for Simulation 3 (Figure 8). These are different ways of visualising how the model perceives the similarity between items throughout the course of learning. Early in training, the network groups items essentially at random, since the weights are initialised to very small random values. Later in training, the network's representations capture both the surface similarities and the relational similarities between items (in contrast to the training patterns, depicted in Figure 3). Progressive differentiation in semantic networks has been explored more extensively in previous work (Flusberg et al., 2010; Rogers & McClelland, 2004, 2008).

Several proponents of the *Structural* approach to analogy have suggested that a defining feature of analogical reasoning is the ability to perform *structured pattern completion* (Gentner & Markman, 1993, 1995), which refers to a process whereby 'a partial representation of the target is completed based on its structural similarity to the base' (Gentner & Markman, 1995). This is typically defined in contrast to *simple pattern completion*, which is 'based on the vector similarities of the current activation pattern to previously learned patterns' (Gentner & Markman, 1993). *Emergent* models are frequently criticised for only being able to perform simple pattern completion, which is unable to account for a great deal of sophisticated human behaviour (as we detailed in the introduction; see also Gentner & Markman, 1995). As we have shown, however, our model behaves in a way that is perfectly captured by the concept of structured pattern completion: it draws inferences based on information from patterns that do not share strong (or, in some cases, any) similarity in terms of raw vector overlap. Rather, the patterns that come to most strongly influence the model's inferences about novel inputs are those that share the most similar structural relationships with the other training items.

It is important to clarify what aspects of the environment we believe are encoded in our training patterns. Many of these patterns, such as those representing the visual features of the family members, might be thought of as arising from perception. However, others, particularly those representing familial relations such as 'mother_of' and 'owner_of', are much more likely to be encoded linguistically than visually. That is, part of our story is that learners hear language describing the people and things around them at the same time as they experience them perceptually, and these different sources of information are integrated whenever (as we think is almost always the case) there is some coherent covariation of information between the several sources (Rogers & McClelland, 2004).

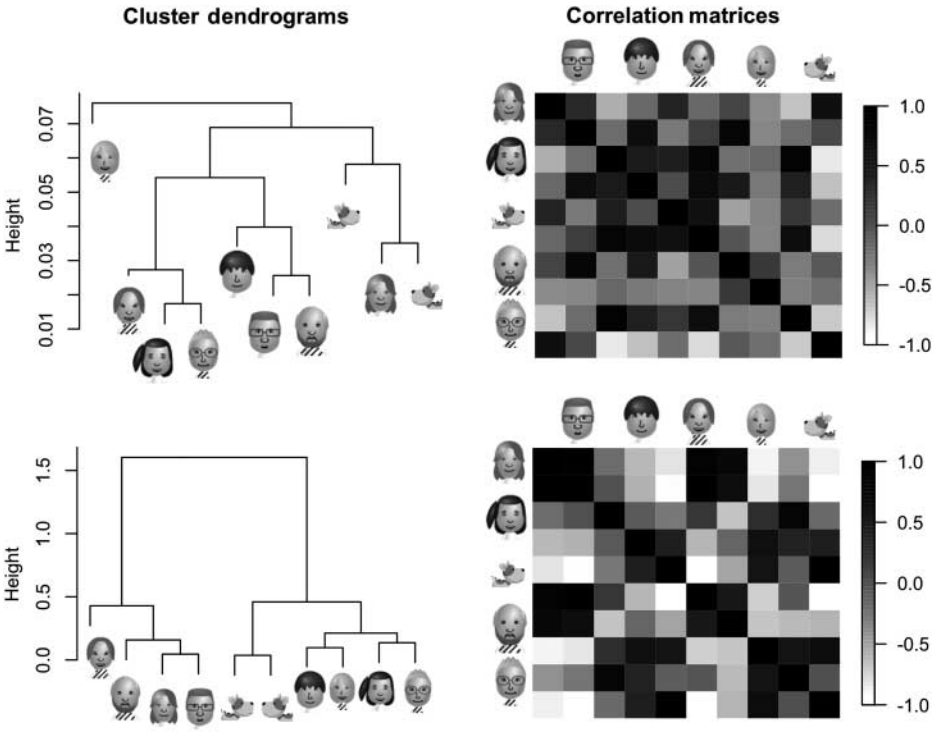

Figure 8. Each of the plots above illustrates the similarity structure of the learned Subject Representations in Simulation 3. Hierarchical clusters are on the left and correlation matrices are on the right. Early in training (the upper panels), the network does not group individuals by family or relation. Later in training, at 1300 epochs (the lower panels), the network has aligned the families according to their relational similarity.

This is consistent with a great deal of empirical work demonstrating that relational language facilitates analogical inference and helps drive the relational shift in analogical development (Gentner, Simms, & Flusberg, 2009; Loewenstein & Gentner, 2005; Rattermann & Gentner, 1998). It is also consistent with other research on how language affects semantic processing more generally, both in experimental studies (Boroditsky, 2001; Fausey, Long, Inamori, & Boroditsky, 2010; Lupyan, Rakison, & McClelland, 2007) and computational models (Andrews, Vigliocco, & Vinson, 2009; Dilkina, McClelland, & Boroditsky, 2007; Flusberg et al., 2010). This approach views relational labels as another set of environmental regularities, serving the function of augmenting the statistical structure of the environment in ways that facilitate learning analogical representations, rather than as explicitly symbolic representations in the brain.

One of the major advantages of this approach is that it allows us to address how analogy naturally emerges over the course of development without having to posit additional, complex cognitive machinery. The Rumelhart model and its descendants (Rogers & McClelland, 2004, 2008) have been used to address a diverse set of findings in conceptual development that had previously been thought to require more complex explanations based on explicitly structured representations or innate, domain-specific constraints (e.g. Carey, 2009; Keil, 1992; Murphy & Medin, 1985). These include basic-level effects in categorisation (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) and illusory correlations of perceptual features in categorised objects (Gelman, 1990), among others. In a similar vein, our extension provides an alternative way of explaining the shift that occurs as children begin to make use of relational information in making inferences (Gentner & Ratterman, 1991). The classic, knowledge-based account of this shift suggests that as children's knowledge of relational structure grows (partially through exposure to relational

language), they begin to use this structure to make inferences via a distinct mapping or alignment process (Gentner, 1988; Gentner & Ratterman, 1991; Rattermann & Gentner, 1998). We agree with this account to the extent that in our model, the relational shift takes place as sensitivity to relational information increases (as opposed to an innate bias towards using relational structure that is initially dominated by perceptual information; cf. Goswami, 1992). However, our model exhibits this shift without proposing explicitly structured knowledge and additional mechanisms for mapping across these structures. This is a parsimonious account that is more easily integrated with what we know about the ontogeny of complex cognitive systems (Elman et al., 1996; Spencer, Thomas, & McClelland, 2009).

In addition, our approach is well suited to explain why children do not always generalise based on relational similarity. When perceptual information, rather than relational information, is more likely to yield correct inferences, people are more likely to rely on perceptual information (Bulloch & Opfer, 2009). Our model simply learns what kind of information is most likely to yield the best inferences and represents this knowledge in a way that supports generalisation. In the simulations we present, relying on relational information yields the best predictions; however, as others have shown (e.g. Rogers & McClelland, 2004), the Rumelhart model can also leverage featural similarity. In this way, our model is well suited to explore context sensitivity in a way that 'mapping' algorithms like SME and LISA are not.

4. General discussion

In this paper, we have grounded analogical reasoning in a sub-symbolic account of learning general semantics in an attempt to provide an *Emergent* approach to analogical inference. This may seem surprising, as analogical inference has been thought to require explicit symbolic representations as well as a distinct mapping mechanism operating over those symbols (Gentner, 2010a; Gentner & Markman, 1993, 1995; Holyoak & Hummel, 2000; Hummel, 2010; Markman, 1999). In this general discussion, we would like to address several sources of tension between our research program and the more dominant *Structural* approach to understanding the nature of analogy.

As we pointed out in the introduction, the persuasiveness of our claims depends in part on whether the tasks we modelled should really be considered 'analogical'. After all, if we have only succeeded in capturing some generic associative learning task then we really have not made any novel contributions to our understanding of the nature of analogy. However, there are several compelling reasons for accepting the idea that our model is genuinely performing analogical inference. We asked the model to make an inference about a pattern it was not trained on, and to succeed at this task the model had to draw on information it had already learned about another family that shared relational structure with the target family. Furthermore, it had to leverage this shared relational structure in the face of shared surface feature similarity between the two families. This ability to map a relation from one domain to another, sometimes called *copying with substitution*, has been singled out by researchers as a hallmark of mature analogical reasoning (Gentner, 1983, 1989, 2003; Gentner & Markman, 1995; Holyoak et al., 1994; Hummel & Holyoak, 2003). As we have already pointed out, this sort of behaviour is well captured by the concept of *structured pattern completion* (Gentner & Markman 1993, 1995).

Furthermore, our task bears strong similarities to behavioural tasks used by other analogy researchers, such as transitive inference tasks, problem analogies, and class-inclusion tasks. For instance, transitive inference tasks require that people abstract over superficial similarities between conceptual domains and, instead, focus on relational similarity (e.g. Brown, 1989; Gentner, 1989; Gentner & Loewenstein, 2002; Gentner & Ratterman, 1991; Goswami, 1995; Goswami & Pauen, 2005; Loewenstein & Gentner, 2005; Rattermann & Gentner, 1998; for a connectionist account of the transitive inference task, see Kumaran & McClelland, 2012). These tasks appear to draw

on many of the same underlying cognitive mechanisms as our own task. Indeed, previous work has shown that young children show improvement in their ability to perform transitive inference when they practice with family trees (Mutafchieva & Kokinov, 2007).

On the other hand, if analogical reasoning is defined not in terms of behaviour or task demands, but rather as the subset of tasks that *require* explicit mapping over structured symbolic representations at the algorithmic level, then the natural conclusion is that whatever our model is doing is not analogy. However, if this were the case, then previous studies that have made use of tasks structurally similar to our own would likewise not count as studies of analogy (e.g. Brown, 1989; Rattermann & Gentner, 1998). Moreover, it would no longer be theoretically useful to talk about structured pattern completion as a unique or distinct property of analogical inference. Rather, it would have to be seen as a more general property of cognitive systems that can be implemented in a variety of formal frameworks. Therefore, even if the conclusion one wants to draw is that our model is not performing analogical inference, this would still advance our understanding of the very nature of analogy by shedding light on what tasks may or may not require *true* analogical inference. We suggest, then, that even the most ardent proponents of the *Structural* approach can gain new insights from the work presented here.

Clearly, our preferred interpretation is to conclude that the model is genuinely performing analogical inference. In fact, we would go so far as to suggest that the model is actually implementing mapping and other processes that have typically been carried out by specialised mechanisms, but doing so in an emergent and approximate fashion. On this view, processes like mapping may be thought of as a useful computational-level description (Marr, 1982) of a system that carries out analogical reasoning (though it may not be the only possible description of such a system; e.g. Leech et al., 2008). Therefore, we would agree that the optimal function or goal of any analogy model might be to map relational structure between domains and to use this structural alignment to guide inference. However, we do not think that this *necessarily* has any direct implications for what sorts of algorithms are used to instantiate this function. Instead, in the *Emergent* approach we separate these levels of description and explore the ways in which lower level algorithmic mechanisms like error-driven learning can give rise to complex, organised behaviour over the course of development (see also Elman et al., 1996; Rogers & McClelland, 2004).

As we have alluded to already, there are several strengths of this approach. First, it offers a parsimonious account of many behavioural findings because it does not require positing additional complex cognitive machinery (Rogers & McClelland, 2004). The Rumelhart network and related connectionist architectures have been shown to capture and help explain a variety of phenomena – the developmental trajectory of semantic learning (e.g. Rogers & McClelland, 2004; Schapiro & McClelland, 2009), peoples' sensitivity to the structure of experience (e.g. Cohen, Dunbar, & McClelland, 1990; Elman, 1990), and the degradation of semantic knowledge as a result of semantic dementia (e.g. Dilkina, McClelland, & Plaut, 2008). We have extended the scope of the Rumelhart network to simple relational reasoning without adding a mechanism to explicitly map representations or bind roles and fillers – mechanisms at the heart of *Structural* models of analogical reasoning.

Second, it allows us to naturally address the development of cognitive functioning and to link this up with what we know about biological and neural development more generally (Elman et al., 1996; Leech et al., 2008; Spencer, Thomas, & McClelland, 2009). In fact, *Structural* models of analogy have been criticised for failing to account for how the mechanisms they instantiate could possibly develop in the first place (Leech et al., 2008). The DORA model (Discovery of Relations by Analogy) (Doumas et al., 2008) attempts to address this concern in the *Structural* framework. However, we see several important differences between DORA and *Emergent* models, particularly with respect to the mechanisms that are thought to be necessary for analogical reasoning. Proponents of DORA argue:

> Discovering a relation and representing it in a form that can support relational thinking entail solving three problems. First, there must be some basic featural invariants that remain constant across instances of the relation, and the perceptual/cognitive system must be able to detect them. Second, the architecture must be able to isolate these invariants from the other properties of the objects engaged in the relation to be learned. Third, it must be able to predicate the relational properties – that is, represent them as explicit entities that can be bound to arbitrary, novel arguments. (Doumas et al., 2008, p. 2)

As we argue throughout this paper, we disagree with this view of what relational reasoning entails. Our model does not require the explicit (symbolic) representation of the arguments and does not view relations as operators over these symbolic representations (see Leech et al., 2008, for related additional concerns with DORA).

While some proponents of the *Structural* approach have argued that the algorithms supporting analogical inference should directly reflect the computational description of mapping over explicitly structured representations (Doumas et al., 2008, Holyoak and Hummel, 2008; Hummel, 2011; Kemp & Tenenbaum, 2008; Marcus, 2001; Marcus & Keil, 2008; Opfer & Doumas, 2008; but cf. Marr, 1982), in principle, it is not necessary for representations to be *explicitly* structured in order to approximate processes like alignment and projection (Chalmers, French, & Hofstadter, 1992; see also Frank et al., 2009; Monner & Reggia, 2012). Other researchers have repeatedly exhibited emergent connectionist models that have compositional structure or that can perform comparable tasks (Chalmers, 1990; Elman, 1990; Leech et al., 2008; Smolensky & Legendre, 2006). (For an early theoretical discussion of this issue, see van Gelder, 1990.)

In our *Emergent* approach, relational reasoning is natural by-product of learned, distributed representations. The use of distributed and graded representations more accurately reflects the graded and quasi-regular nature of both the environment we live in and human behaviour in general (Rogers & McClelland, 2008; Spivey, 2007). Even tasks that appear to require discrete responses are often supported by more graded and dynamic cognitive mechanisms (Spivey, 2007).

Further, the principles that we have documented here have been shown to underlie recurrent networks that are capable of learning more complex relationships (e.g. hierarchical role-filler bindings; McClelland & Kawamoto, 1986; St. John & McClelland, 1990). As a result, we are confident that our approach is capable of scaling up to even more complex and hierarchical structures. For instance, a recurrent model would likely be able to learn that (a) *John loves Mary*, *Mary loves Bill*, and *John is jealous of Bill*; (b) *Sally loves Fred* and *Fred loves Tina*; and, in turn, to infer (c) *Sally is jealous of Tina*. However, as currently configured, our model is not capable of representing this kind of hierarchical information. Nevertheless, we believe this kind of inference to be within the scope of the *Emergent* framework and are eager to explore this hypothesis in future work. Again, we wish to emphasise that this is ultimately an empirical question rather than a logical one.

This is not to say that our approach can account for all human analogical reasoning. Far from it! As we described in our final series of simulations, we have only attempted to capture the sort of analogical inference that emerges as a result of a gradual learning process over the course of development. This may include any task that requires making inferences based on shared relational structure that has been stored in long-term memory, including phenomena like reasoning with conceptual metaphors (Flusberg et al., 2010; Thibodeau et al., 2009). Other researchers have captured additional aspects of analogical reasoning in theoretically related *Emergent* models. For example, Leech et al. (2008) used a recurrent connectionist architecture to simulate analogical reasoning in tasks that take the classic 'A:B::C:D' form. However, even the approach of Leech et al. (2008) may not be able to represent some kinds of analogy problems, in which highly structured knowledge is learned and leveraged for inference extremely quickly (see commentary responses in Leech et al., 2008).

Many of the analogy tasks used in previous work, which models like SME and LISA can capture so well, seem to rely on cognitive processes which we did not even attempt to simulate (Bowdle &

Gentner, 1997; Clement & Gentner, 1991; Markman & Gentner, 1997; Morrison et al., 2004). For example, we would agree that some analogy tasks may rely on working memory and cognitive control processes, as well as one-shot learning and episodic memory (e.g. Gick & Holyoak, 1980). Though we have highlighted the fact that our model performs structured pattern completion, this is not the only defining feature of analogy (Gentner & Markman, 1993, 1995). For instance, mature analogical reasoning is highly flexible and allows for multiple interpretations of a single item in different comparisons, as well as multiple interpretations of any given comparison (Gentner & Markman, 1995). This sort of flexibility lies outside the scope of our particular model and the task we set out to simulate, and it may in fact require additional, online cognitive capacities like working memory (Morrison et al., 2004).

Moreover, some aspects of analogical reasoning may demand much richer linguistic abilities than we implemented in our model. As noted above, we treat relational language as an environmental cue, encoding a certain kind of statistical structure that is then used to shape semantic representations. While this is one important role of language in analogical reasoning, it may not be the only one; the ability to verbally re-describe a situation to oneself, for example, is an important tool in many higher level reasoning tasks (Williams & Lombrozo, 2010).

Further, despite our best efforts to systematically vary the training and test conditions of the model, the simplicity of the target domains may be interpreted as a drawback of our approach. We see a trade-off between maximising external validity – the degree to which the model reflects the truly complex structure of the real world – and accessibility – our primary goal of presenting as clearly as possible how a set of well-understood, domain-general, and neurally inspired mechanisms can support high-level relational inference. Balancing these goals is a challenge, particularly to theoretical modelling in cognitive science (McClelland, 2009).

In principle, we think that the model could build representations of more complex domains that contain many more relations and features without losing the ability to extract relational structure and leverage it in a flexible and powerful way. However, as the complexity of the source and target domains (and other domains that may be irrelevant to the mapping) increases, the model may have a more difficult time making some analogical inferences. We do not view this as a flaw of the model though. People too have a difficult time solving problems by analogy and selecting relations from memory. Gick and Holyoak (1980), for instance, found that only a very small percentage of people solved a target problem spontaneously by analogy. Performance improved when people were given a hint to use an analogy or primed with multiple source domains. Our model is consistent with this idea, because in some sense it takes 'activating' a relation (i.e. giving the model a hint) to get it to complete a relational inference. That is, our model offers one account for why giving a hint is helpful for solving problems by analogy: it leads to a process of structured pattern completion that helps to select the appropriate relation and solve the target problem.

We emphasise that this is only a first step in a process of using emergent models to help elucidate the mechanisms that underlie analogical reasoning. In the future, we plan to build on the current work by exploring when and how emergent models can abstract and leverage relational structure in more complex domains, with more variability between the source and target structures, and when these structures encode a greater number of relations.

In sum, these issues point to a possible heterogeneous view of analogy. While some forms of analogical inference may naturally emerge over the course of development due to the operation of low-level learning mechanisms, other forms of analogical reasoning may only be possible by exercising additional processes like cognitive control and working memory. Therefore, we would like to suggest that one major unsolved problem in the field is the integration of the kind of slow-learning, semantic cognition model described in this paper with the online, structurally explicit models already in place. The extensive and valuable work on models such as SME and LISA over the past 20 years, no less than the connectionist models we have implemented, must be used to guide future research into analogical processing across development, in behaviour and in the brain.

In this sense, we can draw an analogy between this debate and the *past tense debate* (McClelland & Patterson, 2002a, 200b; Pinker & Ullman, 2002a, 2002b), in which competing mechanisms were offered to explain how people generate past-tense inflections of verbs. One view, the words and rules approach, was structural and appealed to a complimentary set of mechanisms – a lexicon and a rule-governed syntactic system (Pinker & Ullman, 2002a, 2002b). The other, emergent view appealed to a single, distributed system that was able to learn and leverage knowledge of clusters of quasi-regular verbs (McClelland & Patterson, 2002a, 200b). As is the case here, the two models of past-tense inflection are difficult to differentiate in behavioural experiments, have contrasting strengths and weaknesses, and maintain supporters and detractors to this day. Nonetheless, the debate itself was the impetus for a great deal of research that has only deepened our understanding of language processing. Our hope is that the simulations and argument we have presented will similarly encourage people to think critically about the mechanisms that support analogy and motivate continued theoretical and empirical work on analogical reasoning.

5. Conclusion

We have shown that analogical mapping and inference can emerge over the course of development in a distributed model based on a simple, domain-general learning mechanism. Beyond this, our results suggest that analogy may indeed lie at the very core of cognition, but for reasons quite different from those suggested by previous researchers. Our framework suggests that distributed representations that support analogical inference may arise naturally, spontaneously, and pervasively throughout development wherever there is shared relational structure between domains in the environment. On this account, which we have described in terms of structured pattern completion, our experience of any piece of the world may be vastly enriched by all our other analogically relevant experiences. This could allow us to impute structure in novel domains even with minimal exposure, to map higher order relationships learned through language onto both concrete and abstract domains, and even to correlate the structure of language with the structure of the world – in short, many of the very things that seem to make us smart.

Acknowledgements

The authors would like to thank Jay McClelland and Lera Boroditsky for inspiration and helpful, critical discussion of the content of this paper and the issues it addresses. In addition, we would like to thank John Hummel, Alex Doumas, Chris Eliasmith, Dedre Gentner, and Kieth Holyoak for their constructive criticism and helpful suggestions on previous iterations of the argument and simulations we have presented here. This material is based on work supported under a National Science Foundation Graduate Research Fellowship and a Stanford Graduate Fellowship.

Notes

1. It should be noted that this is not the *only* way to sub-divide analogy. For example, Hall (1989) suggests that analogy consists of Recognition, Elaboration, Evaluation & Transfer, and Consolidation. However, there is a great deal of overlap between the ways in which different researchers conceive of these sub-processes (e.g. Elaboration is very similar to Mapping).
2. For convenience, we often use labelled or even localist input and output units. We interpret these patterns as observed states of the world, including linguistic and non-linguistic perceptual-motor experience, which are used both to predict future states and to provide corrective feedback to these predictions (Flusberg et al., 2010; Rogers & McClelland, 2008). The input and output patters can be viewed as symbolic. However, unlike traditional symbolic models of cognition, our model does not perform operations over these symbols, but over the learned re-representations of the symbols in the internal layers. What is important for the internal layers is the statistical structure encoded by the entire set of input and output vectors.
3. It bears mentioning that an argument for the existence of slow- and fast-learning systems was not the only lesson of McClelland et al. (1995). In fact, one of their major goals in positing these complementary systems was to solve

the difficult problem of how new information could be integrated with previously learned information without overwriting that previous information (i.e. catastrophic interference).

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*, 463–498.

Blanchette, I., & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory & Cognition, 28*, 108–124.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive psychology*, *43*(1), 1–22.

Bowdle, B., & Gentner, D. (1997). Informativity and asymmetry in comparisons. *Cognitive Psychology*, *34*(3), 244–286.

Brown, A. L. (1989). Analogical learning and transfer: What develops? In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 369–412). New York: Cambridge University Press.

Brown, T. L. (2003). *Making truth: Metaphor in science*. Urbana-Champaign, IL: University of Illinois Press.

Bulloch, M. J., & Opfer, J. E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science, 12*, 114–122.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.

Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science*, *2*(1), 53–62.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence, 4*, 185–211.

Chen, Z. (1996). Children's analogical problem solving: The effects of superficial, structural, and procedural similarity. *Journal of Experimental Child Psychology, 62*, 410–431.

Clement, C., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, *15*(1), 89–132.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel-distributed processing account of the Stroop effect. *Psychological Review, 97*, 332–361.

Cohen, N. J., & Eichenbaum, H. E. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Dilkina, K., McClelland, J. L., & Boroditsky, L. (2007). How language affects thought in a connectionist model. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 215–220). Austin, TX: Cognitive Science Society.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology, 25*, 136–164.

Doumas, L., Hummel, J., & Sandhofer, C. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 365–395). Cambridge, MA: MIT Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & S. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery and change* (pp. 461–493). Washington, DC: APA Press.

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience, 1*, 41–50.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Fausey, C. M., Long, B. L., Inamori, A., & Boroditsky, L. (2010). Constructing agency: The role of language. *Frontiers in Psychology*, *1*(0), 1–11.

Flusberg, S. J., Thibodeau, P. H., Sternberg, D. A., & Glick, J. J. (2010). A connectionist approach to embodied conceptual metaphor. *Frontiers in Psychology*, *1*(0), 1–11.

Forbus, K., Gentner, D., & Law (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19*, 141–205.

Frank, S. L., Haselager, W. F. G., & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition, 110*, 358–379.

French, R. (1995). *The subtlety of sameness: A theory and computer model of analogy-making.* Cambridge, MA: MIT Press.

French, R. (2002). The computational modeling of analogy-making. *Trends in Cognitive Sciences*, *6*(5), 200–205.

van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science, 14*(3), 355–384.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science*, *14*(1), 79–106.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, *59,* 47–59.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). London: Cambridge University Press (Reprinted in Knowledge acquisition and learning, 1993, 673–694).

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, MA: MIT Press.
Gentner, D. (2010a). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752–775.
Gentner, D. (2010b). Psychology in cognitive science: 1978–2038. *Topics in Cognitive Science*, *2*(3), 328–344.
Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *WIREs Cognitive Science*, *2*, 266–276.
Gentner, D., & Loewenstein, J. (2002). *Learning: analogical reasoning. Encyclopedia of Education* (2nd ed.). New York, NY: Macmillan.
Gentner, D., & Markman, A. B. (1993). Analogy – watershed or waterloo? Structural alignment and the development of connectionist models of cognition. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5 [NIPS Conference]* (pp. 855–862). San Francisco, CA: Morgan Kaufmann.
Gentner, D., & Markman, A. B. (1995). Analogy-based reasoning in connectionism. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 91–93). Cambridge, MA: MIT Press.
Gentner, D., & Ratterman, M. (1991). Language and the career of similarity. In S. Gelman & J. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225–277). London: Cambridge University Press.
Gentner, D., Simms, N., & Flusberg, S. (2009). Relational language helps children reason analogically. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th annual conference of the cognitive science society* (pp. 1054–1059). Austin, TX: Cognitive Science Society.
Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355.
Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Erlbaum.
Goswami, U. (1995). Transitive relational mappings in three and four year olds: The analogy of Goldilocks and the three bears. *Child Development, 66*, 877–892.
Goswami, U., & Pauen, S. (2005). The effects of a 'family' analogy on class inclusion reasoning by young children. *Swiss Journal of Psychology, 64*, 115–124.
Hall, R. P. (1989). Computational approaches to analogical reasoning: A comparative analysis. *Artificial Intelligence, 39*, 39–120.
Hinton, G. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society* (Vol. 1, pp. 1–12). Hillsdale, NJ: Erlbaum.
Hofstadter, D. (1996). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* New York, NY: Basic Books.
Hofstadter, D. (2001). Analogy as the core of cognition. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 499–538). Cambridge, MA: MIT Press.
Hofstadter, D. R., & Mitchell, M. (1994). The Copycat Project: A model of mental fluidity and analogy-making. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol. 2: Analogical connections* (pp. 31–112). Norwood, NJ: Ablex.
Holyoak, K. J., Gentner, D., & Kokinov, B. (2001). The place of analogy in cognition. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 1–19). Cambridge, MA: MIT Press.
Holyoak, K. J., & Hummel, J. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 229–264). Hillsdale, NJ: Lawrence Erlbaum Associates.
Holyoak, K. J., & Hummel, J. E. (2008). No way to start a space program: Associationism as a launch pad for analogical reasoning. *Behavioral and Brain Sciences, 31*, 388–389.
Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15*, 332–340.
Holyoak, K. J., Novick, L. R., & Melz, E. R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. J. Holyoak & J. A. Barnden (Eds.), *Analogical connections: Advances in connectionist and neural computation theory* (Vol. 2, pp. 113–180). Westport, CT: Ablex.
Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.
Hummel, J. E. (2010). Symbolic versus associative learning. *Cognitive Science*, *34*(6), 958–965.
Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*(2), 109–118.
Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.
Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264.
Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture. *Current Directions in Psychological Science*, *14*(3), 153–157.
Keil, F. (1992). *Concepts, kinds, and cognitive development*. Cambridge, MA: The MIT Press.
Kemp, C., & Tenenbaum, J. B. (2008). Structured models of cognition. *Behavioral and Brain Sciences, 31*, 717–718.
Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review, 119*, 573–616.
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.
Leech, R., Mareschal, D., & Cooper, R. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, *31*, 357–378.
Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*, 315–353.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science, 18*, 1077–1083.
Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.
Marcus, G. F., & Keil, F. C. (2008). Concepts, correlations, and some challenges for connectionist cognition. *Behavioral and Brain Sciences, 31*, 722–723.
Markman, A., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, *8*, 363–367.
Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Lawrence Erlbaum Associates.
Markman, A. B., & Laux, J. P. (2008). Analogical inferences are central to analogy. Commentary on Leech, Mareschal & Cooper. *Behavioral and Brain Sciences*, *31*, 390–391.
Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.
McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1*, 11–38.
McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science, 2*, 751–770.
McClelland, J. L., & Goddard, N. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus, 6*, 654–665.
McClelland, J. L., & Kawamoto, A. H. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In J. L. McClelland, D. E. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition 2: Applications* (pp. 318–362). Cambridge, MA: MIT Press.
McClelland, J. L., McNaughton, B. L., & 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*, 419–37.
McClelland, J. L., & Patterson, K. (2002a). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6*, 465–472.
McClelland, J. L., & Patterson, K. (2002b). 'Words *Or* Rules' cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences, 6*, 464–465.
McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*, 310–322.
Mitchell, M. (1993). *Analogy-making as perception: A computer model.* Cambridge, MA: MIT Press.
Monner, D., & Reggia, J. A. (2012). Emergent latent symbols in recurrent neural networks. *Connection Science, 24*, 193–225.
Morrison, R., Krawczyk, D., Holyoak, K., Hummel, J., Chow, T., Miller, B., & Knowlton, B. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*(2), 260–271.
Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.
Mutafchieva, M., & Kokinov, B. (2007). Does the family analogy help young children to do relational mapping. In *Proceedings of the European cognitive science conference*. Hillsdale, NJ: Erlbaum.
Nersessian (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science* (Vol. XV). Minneapolis: University of Minnesota 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*, 611–646.
Opfer, J. E., & Doumas, L. A. A. (2008). Analogy and conceptual change in childhood. *Behavioral and Brain Sciences, 31*, 723–724.
O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review, 108*, 311–345.
Penn, D., Holyoak, K., & Povinelli, D. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(02), 109–130.
Pinker, S., & Ullman, M. T. (2002a). Combination and structure, not gradedness, is the issue. *Trends in Cognitive Sciences, 6*, 472–474.
Pinker, S., & Ullman, M. T. (2002b). The past and future of the past tense. *Trends in Cognitive Sciences, 6*, 456–463.
Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*, 909–957.
Rattermann, M., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development*, *13*, 453–478.
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition*. Cambridge, MA: MIT Press.
Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*, 689–749.
Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.
Rumelhart, D. (1990). Brain style computation: Learning and generalization. In S. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405–420). San Diego, CA: Academic Press.
Rumelhart, D., & Todd, P. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.
Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition, 110*, 395–411.
Smolensky, P., & Legendre, G. (2006). *The harmonic mind*. Cambridge, MA: MIT Press.

Spencer, J. P., Thomas, M. S., & McClelland, J. L. (Eds.). (2009). *Toward a new grand theory of development? Connectionism and dynamic systems theory re-considered*. New York, NY: Oxford University Press.

Spivey, J. (2007). *The continuity of mind*. Oxford: Oxford University Press.

St. John, M. F. (1992). The story gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science, 16*, 271–306.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217–257.

Thelen, E., & Smith, L. B. (1998). Dynamic systems theories. In R. M. Lerner (Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 563–635). New York, NY: Wiley.

Thibodeau, P., McClelland, J. L., & Boroditsky, L. (2009). When a bad metaphor may not be a victimless crime: The role of metaphor in social policy. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 809–814). Austin, TX: Cognitive Science Society.

Welling, H. (2007). Four mental operations in creative cognition: The importance of abstraction. *Creativity Research Journal, 19*, 163–177.

Williams, J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.