TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 9 (2017) 694–718

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12277

This article is part of the topic "2016 Rumelhart Prize Issue Honoring Dedre Gentner," Jeffrey Loewenstein and Arthur B. Markman (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/doi/10.1111/tops.2017.9.issue-2/issuetoc

Representation and Computation in Cognitive Models

Kenneth D. Forbus, Chen Liang, Irina Rabkina

Department of Computer Science, Northwestern University

Received 19 May 2016; received in revised form 26 January 2017; accepted 17 March 2017

Abstract

One of the central issues in cognitive science is the nature of human representations. We argue that symbolic representations are essential for capturing human cognitive capabilities. We start by examining some common misconceptions found in discussions of representations and models. Next we examine evidence that symbolic representations are essential for capturing human cognitive capabilities, drawing on the analogy literature. Then we examine fundamental limitations of feature vectors and other distributed representations that, despite their recent successes on various practical problems, suggest that they are insufficient to capture many aspects of human cognition. After that, we describe the implications for cognitive architecture of our view that analogy is central, and we speculate on roles for hybrid approaches. We close with an analogy that might help bridge the gap.

Keywords: Analogy; Representation; Machine learning; Computational modeling; Learning; Relational representations; Symbolic modeling

1. Introduction

What sort of representations does human cognition require? Many proposals have been made (Dietrich & Markman, 2000; Markman, 1998), but recently distributed representations have become extremely popular. A distributed representation describes examples and concepts in terms of a set of numbers, with the common metaphor being patterns of activation in neural systems. Mathematically, researchers use vectors, matrices, or tensors. Distributed representations have some desirable properties: They are relatively easy to

Correspondence should be sent to Kenneth D. Forbus, Computer Science, Northwestern University, 2133 Sheridan Road, Evanston, IL 60201. E-mail: forbus@northwestern.edu

construct from widely available data, the mathematics underlying them is well understood, and today's GPUs provide efficient ways to implement large-scale models. Moreover, they have been used successfully in a variety of commercially important applications, such as speech recognition and machine translation. These successes have caused some to argue that such non-symbolic distributed representations are sufficient to explain human cognition (LeCun, Bengio, & Hinton, 2015; Rogers & McClelland, 2004). We do not believe that this is the case. We believe that symbolic representations—structured descriptions, which include relations and their bindings—are essential for carrying out higher order human cognition. This paper makes that case.

The argument goes as follows:

1. First, we briefly examine some common misconceptions concerning representations and learning in both people and machines, to set the stage.
2. Second, we argue that structured, relational representations are necessary to explain human cognition by drawing on the analogy literature, both direct psychological evidence and results from computational models of analogy.
3. Third, we argue that distributed representations are insufficient for explaining human cognition, using a combination of theoretical arguments and empirical evidence.
4. Fourth, we step back and consider the implications of analogical processing for cognitive architecture, including the prospect of hybrid symbolic/distributed systems.

We close with additional thoughts on open questions involving representation in cognition, and an analogy.

2. Setting the stage

There is a set of long-standing misconceptions that need to be cleared out of the way first, so that we can focus on the matter at hand.

2.1. Misconception: Symbolic means serial, logical, and non-numerical

Symbolic models have integrated structural and numerical information from the start of cognitive science. For example, semantic networks used relationships between nodes, while at the same time, using parallel spreading activation (Collins & Quillian, 1970; Collins & Loftus, 1975). Models of human spatial reasoning typically combine both quantitative and qualitative aspects (e.g., Forbus, 1980; Forbus, Nielsen, & Faltings, 1991; Kuipers, 2000), which correspond to coordinate and categorical aspects of spatial models in the psychological literature (e.g., Holden, Newcombe, & Shipley, 2015; Huttenlocher, Hedges, & Duncan, 1991). Hinton's (1979) classic model of mental imagery is another example of structured representations combined with numerical properties. Both ACT-R (Anderson, 2009) and SOAR (Laird, 2012) use numerical components in their representations, for example, statistical metadata on recency, frequency, and utility for symbolic structures. There are currently

several theoretical frameworks that tightly integrate logic and probability, including Markov Logic Networks (Domingos, Kok, Poon, Richardson, & Singla, 2006) and Bayesian Logic (Milch et al., 2007), while Rosenbloom's (2010) SIGMA cognitive architecture is exploring how to build a complete cognitive architecture, including both symbolic and statistical reasoning, out of graphical models.

IBM's Watson provides an example of how symbolic systems can be parallel, incorporate non-logical techniques, and rely on probabilities. In 2011, it demonstrated a revolutionary improvement in what is known as "factoid" question answering, that is, answering factual questions by retrieving and combining information from textual sources. Accuracy in factoid Q/A performance using purely statistical techniques had hovered around 30% for years. The IBM team was the first to achieve 85–90% accuracy (Ferrucci, 2012). They parsed questions into structured descriptions, and then used a combination of statistical language-level methods and structured inference techniques to explore, in parallel, many possible answers. Classifiers learned during training were used to estimate confidence in intermediate results throughout processing. This combination led to performance that was both broad enough and rapid enough to beat some of the top human Jeopardy! players in the world, in a game that requires real-time competitive natural language question answering.

2.2. Misconception: Symbolic systems don't do well at learning

Again, one counterexample among many is IBM's Watson. Its performance rests in part on the PRISMATIC knowledge base (Fan, Kalyanpur, Gondek, & Ferrucci, 2012). PRISMATIC provides shallow structured knowledge that includes probabilities, which are used to help determine the likelihood of answers. PRISMATIC was learned by reading massive amounts of English text (Wikipedia, several encyclopedias, literary works, and many other documents). The reading process involved constructing full syntactic parses and named entity recognition over the entire corpus, constructing over 900 million syntactic frames. The parser used for this was IBM's English Slot Grammar parser (McCord, Murdock, & Boguraev, 2012), which was also used in understanding questions in the real-time system. Importantly, unlike other parsers, it uses a hierarchical conceptual ontology of 160 concepts in its grammar rules.

In artificial intelligence, work on statistical relational learning abounds,[1] including a thriving inductive logic programming community.[2] In cognitive science, many other symbolic models of learning have been built; for example, both SOAR and ACT-R have been used to model many learning tasks. Other examples include explanation-based learning systems (e.g., DeJong, 1993; Minton et al., 1989; Mitchell, Keller, & Kedar-Cabelli, 1986) and systems that learn problem solving (e.g., VanLehn & Jones, 1991). We discuss analogical models of learning below.

2.3. Misconception: Artificial neural networks are biologically plausible

An excellent summary of six objections from neuroscience regarding the biological plausibility of deep learning can be found in Bengio, Lee, Bornschein, and Lin (2015).

Some of these objections include that real neurons communicate via spikes instead of continuous values, and that back-propagation would require "precisely clocked" computation to keep feed-forward and back-propagation operations separate. In addition to these problems, the increasing evidence for the role of glial cells in synaptic firing (e.g., Allen & Barnes, 2005; Auld & Robitaille, 2003; Eroglu & Barres, 2010) suggests that models based solely on neurons and their connections are at best incomplete. By contrast, neuroimaging results from Chatterjee and his collaborators (Chatterjee, 2010; Jamrozik, McQuire, Cardillo, & Chatterjee, 2015) provide neural evidence for relational representation in human cognition.

2.4. Misconception: Artificial neural networks learn in human-like ways

Human learning is often very efficient, as measured by the number of examples needed. For example, in Marcus, Vijayan, Bandi Rao, and Vishton's (1999) studies of rule learning, 7-month-old infants learned a grammar-like pattern after hearing just 16 examples, three times each. An analogy-based model learned a structural pattern within the same number of stimuli as human children, while connectionist models required orders of magnitude more exposures (Kuehne, Gentner, & Forbus, 2000). In contrast, deep learning systems require enormous numbers of examples. For example, AlphaGo beat the world's best Go player, but only after being trained on over 30 million games (Silver et al., 2016)—more than any human could play in a lifetime.[3] We discuss more such unrealistic data requirements below.

2.5. Misconception: Machine learning systems, including deep learning systems, learn without human intervention

Biological organisms are miracles of autonomous learning, compared to the state of the art in today's machine learning systems. Today's machine learning systems require continual human intervention: preparing data, deciding on input representations and hyper-parameters,[4] running the algorithm (or algorithms), inspecting the results, and repeating this cycle many times (Bengio, 2012). When enough resources are available—data and human engineering—systems have sometimes achieved human-level performance on the specific tasks that they are engineered for. AlphaGo, which combined Monte-Carlo tree search with a static evaluator trained via deep learning, was developed by experts who created training regimes based on how the system was evolving (Silver et al., 2016). This training process included setting up representation conventions for the components, including "a small number of handcrafted local features [which] encode common-sense Go rules." Even IBM's Watson, which used both machine learning and symbolic methods, involved roughly 25 researchers working for 4 years, in 2-week cycles of development, testing, and evaluation (Baker, 2011). This in no way detracts from the magnitude of these achievements. Nevertheless, it does indicate that today's learning systems require the guidance of experts who are fully familiar with their internals,[5] unlike people being taught or people learning on their own.

To summarize: Symbolic relational representations have been used in systems that learn, and do not restrict cognitive systems to use only logic, can operate in parallel as well as serial and have proven effective at large-scale learning. Artificial neural networks, despite their practical successes, are not accurate models of biological systems, do not learn as rapidly (in terms of number of examples) as people do, and are hand crafted in their architecture and training for specific tasks. In all of today's learning systems, human experts are heavily involved in their construction and training, where training includes the kind of inspection and manipulation of their internals that is not possible when training people or other animals.

While both symbolic and neural systems have useful properties, many of the supposed limitations of symbolic systems are myths, and many of the supposed advantages of neural systems are myths. The state of the art of learning systems today, in either form, requires far more human intervention, in task-specific ways, than biological systems do.

3. The necessity and effectiveness of structure in human representations

Structured information is a crucial part of human cognition. Causal reasoning, plans, explanations, and discourse are all everyday examples of structured information that people process routinely. Any model that aspires to explain human cognition must be able to provide an account of how we process such information. The simplest account is that we have internal symbolic representations, constructed to capture the connections between different things in the world—their relationships—which are needed to comprehend them. This was one of the original approaches of cognitive science, and it continues to be productively used by many scientists. Indeed, even some non-symbolic approaches end up seeking ways to bridge to, or implement, symbolic representations: Examples include Barsalou's (1999) perceptual symbol systems, neural Turing machines (Graves, Wayne, & Danihelka, 2014), and using temporal binding to implement symbols in distributed systems (Hummel, 2011).

This section argues that symbolic, relational representations are necessary to explain human cognition. The structure of the argument is as follows:

1. Analogy and similarity are central to human cognition.
2. Since structure-mapping theory involves relational representations, including higher-order relations, such representations are necessary for explaining human cognition.
3. Furthermore, this suggests that using symbolic relational representations should provide an effective way to model reasoning and learning across a broad range of cognitive tasks.

We start by reviewing evidence for the importance of structure-mapping in human cognition. We then review how computational models of structure-mapping processes have been used to model a range of cognitive phenomena, including visual problem solving, textbook problem solving (including transfer learning), moral decision-making, and conceptual change. This demonstrates the effectiveness of symbolic modeling for capturing many aspects of human cognition.

3.1. Psychological evidence

Gentner's (1983) structure-mapping theory of analogy proposed that analogy involves constructing correspondences between structured descriptions. This process is often referred to as structural alignment. This theory and related approaches (Doumas, Hummel, & Sandhofer, 2008; Holyoak & Thagard, 1995; Hummel & Holyoak, 1997; Kokinov & French, 2003) have generated considerable empirical support, both in terms of explaining existing findings and predicting new ones (Forbus, 2001). For example,

1. People prefer analogical inferences that are supported by higher order relational structure (Clement & Gentner, 1991).
2. People prefer one-to-one correspondences in analogical mapping (Krawczyk, Holyoak, & Hummel, 2005; Markman, 1997).
3. Structural alignment provides a better model of human similarity judgments than feature-based models (Gentner & Markman, 1995; Markman & Gentner, 1993a), including which direction is preferred in similarity comparisons (Bowdle & Gentner, 1997).
4. Structure-mapping explains several phenomena involving difference detection, including that comparison promotes noticing both commonalities and related differences (Gentner & Gunn, 2001; Markman & Gentner, 1993b); and that it is faster to notice that two items are different when they are very different, but faster to name a difference between them when they are very similar (Sagi, Gentner, & Lovett, 2012).
5. Structure-mapping correctly predicts that metaphor interpretation involves an initial symmetric phase, followed by directional process (Wolff & Gentner, 2011).

There is a substantial body of psychological evidence indicating that analogy and similarity are central to human cognition (Gentner, 2003), and even that analogical processes are instrumental in language learning (Gentner & Namy, 2006; Tomasello, 2003). Furthermore, there is evidence that analogical inferences can occur without conscious awareness (Day & Gentner, 2007; Perrott, Gentner, & Bodenhausen, 2005). If analogy and similarity are central to human cognition, and they rely on structured, relational representations, then that suggests relational representations may be necessary for human cognition. Indeed, it has been argued that our superior relational representation and reasoning capabilities are what set us apart from other primates (Christie, Gentner, Call, & Haun, 2016; Gentner, 2003, 2010; Hofstadter & Sander, 2013; Penn, Holyoak, & Povinelli, 2008).

3.2. Symbolic relational models explain a variety of psychological phenomena

It is one thing to argue that structured descriptions are necessary in human representations; it is another to show that such representations are actually capable of fueling models that reason and learn in human-like ways (Cassimatis, Bello, & Langley, 2008). There is a long history of such models in cognitive science, continuing to this day. For example, cognitive architectures originally aimed at skill learning, such as SOAR (Laird, 2012) and ACT-R (Anderson, 2009), have demonstrated the ability to handle a wider range of

cognitive tasks than any existing machine learning system or system relying on distributed representations. Given the focus of this paper, we do not attempt a complete survey, and instead we highlight some of our computational studies showing the power of structure-mapping.

The large-scale models discussed below use three component models of processes involved in analogy:

1. SME (Falkenhainer, Forbus, & Gentner, 1989; Forbus, Ferguson, Lovett, & Gentner, 2016) models analogical matching. It produces correspondences as well as a numerical similarity score, and *candidate inferences* representing relational structure that can be projected from one description to the other.
2. MAC/FAC (Forbus, Gentner, & Law, 1995) models similarity-based retrieval. Given a probe, it finds an approximation to the most similar case in long-term memory, using a two-phase process.
3. SEQL (Kuehne et al., 2000) and its descendent, SAGE (McLure, Friedman, & Forbus, 2015), model analogical generalization. They operate incrementally, constructing generalizations that preserve the common structure between very close examples. SAGE constructs probabilities for each aligned statement in a generalization, and it can handle disjunctive concepts as well as outliers.

These component models have been incorporated into a cognitive architecture, the Companion cognitive architecture (Forbus, Klenk, & Hinrichs, 2009). This architecture is an exploration of the hypothesis that analogy and qualitative representations are essential to human cognition. We discuss four categories of tasks next to provide evidence for the power of analogical modeling to explain cognitive phenomena.

3.2.1. Analogy in problem solving

We have modeled the use of analogy to import solution plans from previously solved examples (i.e., derivational analogies; Veloso et al., 1995), and it showed that processes promoting expertise in people, such as doing careful qualitative analysis during the solution process, do indeed lead to performance improvements in the system's problem solving (Ouyang & Forbus, 2006). We have also modeled the use of analogy to import principles from prior problems, such as relevant equations (Klenk & Forbus, 2009), a form of transfer learning. In those experiments, a Companion system was trained and tested on a portion of the AP Physics exam by the Educational Testing Service, the organization which creates the test each year. One or two prior examples were sufficient for a Companion to be able to learn transferable knowledge via analogy. Moreover, the operations of the Companion model were compatible with analogy events found by other researchers in protocol studies (Klenk & Forbus, 2007).

Another way that humans use analogy in problem solving is to make comparisons within problems. For example, comparative analysis problems involve sorting a set of scenarios along a dimension (e.g., when trying to pull out a tree stump, which of the four configurations of ropes shown in an accompanying diagram provides the most force?). The Companion architecture has been used to model the solution of such problems from

conceptual physics textbooks (Chang, Wetzel, & Forbus, 2014), using SME to compare relational representations of sketched versions of the diagram (see visual problem solving discussion below). Similarly, the Companion architecture was used to model the solution of problems from the Bennett Mechanical Comprehension test, a widely used test of spatial ability. There, the physical principles were explained to a Companion via examples that integrated visual and conceptual information. New problems were solved, using MAC/FAC to retrieve previous explanations and apply them via analogy to solve new problems (Klenk, Forbus, Tomai, & Kim, 2011).

3.2.2. Analogy in visual problem solving

Most machine learning systems for doing visual tasks, including deep learning systems, start with pixel-level inputs. However, decades of research in vision science indicate that the early stages of visual processing compute edges and other descriptions that combine numerical and structural information (Marr, 1982; Palmer, 1999). We model this by using sketches consisting of digital ink,[6] and using vision techniques to compute relational representations automatically from the ink. This automatic computation is performed via CogSketch (Forbus, Usher, Lovett, Lockwood, & Wetzel, 2011), which is designed as a model of human high-level vision. CogSketch contains multiple levels of visual representation and processing, including gestalt principles for grouping, the ability to decompose ink into edges and recombine it into new shapes, and rudimentary texture detection. This enables CogSketch simulations to automatically construct their own representations and carry out visual re-representation as needed.

One of the key hypotheses underlying CogSketch is that structure-mapping operations play important roles in high-level visual reasoning and learning. Thus, SME, MAC/FAC, and SAGE are all built into CogSketch itself. For example, mental rotation is modeled as a two-pass operation, the first using SME over orientation-independent qualitative representations, and the second being a quantitative computation (Lovett & Forbus, 2013). Analogical generalization (via SAGE) has been used with sketched inputs to learn spatial prepositions (Lockwood, Lovett, & Forbus, 2008) and other kinds of spatial concepts (McLure et al., 2015).

The analogical capabilities of CogSketch have enabled it to be used in modeling several well-known human visual problem-solving tasks. These models have the property that they achieve human-level performance on these tasks, while providing explanations or predictions of human behavior. The models all use the same set of representations, automatically constructed from sketches produced via copy/paste from PowerPoint, with task-specific processing being performed by high-level *spatial routines* (Lovett, 2012). We describe three such tasks next.

The first task is solving geometric analogy problems (Evans, 1968), of the form A is to B as C is to ?, with five options being provided to choose between.[7] The initial CogSketch model produced accurate reaction time predictions based on when the model needed to explore alternate representations (Lovett, Tomai, Forbus, & Usher, 2009). An improved version of the model sheds light on the controversy between two methods proposed in the psychological literature. One method is projection, which is imagining what

difference between A and B would lead to, if applied to C. The other method is second-order comparison, namely comparing the differences between A and B to the differences between C and each of the possible answers. The model shows that projection is more efficient when possible, but that it cannot always be done, so both strategies can be necessary. A model combining both strategies explains the variation in human reaction times with an $R2$ of over 0.76 (Lovett & Forbus, 2012).

The second task is a visual oddity task, which was used to study geometric processing differences between Americans and Mundurukú, an indigenous South American group (Dehaene, Izard, Pica, & Spelke, 2006). Given an array of six images, the goal is to pick the one that is odd or doesn't fit. The CogSketch model used SME to compare the examples in the array to produce generalizations, and to look for an image that was significantly different from generalizations formed from the others. An ablation study on the model was able to suggest reasons for the differences that the original study found between the two groups (Lovett & Forbus, 2011).

The third task is Ravens' Progressive Matrices, a widely used test of human fluid intelligence. Ravens' problems are like geometric analogies, but with one cell of a 2×2 or 3×3 array of images missing, to be filled in by the test-taker. The CogSketch model encodes several strategies argued for in the literature, implemented in terms of analogical comparisons between elements of the array, as well as visual re-representation strategies to reorganize its understanding in order to improve a match. Its performance on the Standard Ravens' test puts it in the 75th percentile, making it better than most adult Americans. Furthermore, it makes reaction-time predictions about human performance that have been confirmed in laboratory experiments (Lovett, Forbus, & Usher, 2010; Lovett & Forbus, 2017).[8]

3.2.3. Analogy in moral decision-making

Human moral judgments can involve both utilitarian factors and sacred values (Baron & Spranca, 1997), values which cannot be violated even if a higher-utility situation would result. For example, if directly killing one person would save many more, people will often not accept doing that, since killing is wrong. The MoralDM model (Dehghani, Tomai, Forbus, & Klenk, 2008) captures the impact of sacred values via a qualitative order of magnitude representation, which makes the utilitarian differences seem negligible in the face of the severity of the violation of such a value. MoralDM uses a combination of rules and analogy to help detect the presence of sacred values in a decision problem. A natural language understanding system (Tomai & Forbus, 2009) is used to construct cases from simplified English stories. Adding new cases improves performance (as measured by matching the majority decisions taken by participants in psychological experiments), and analogical generalization over such cases provides further improvement (Blass & Forbus, 2015).

3.2.4. Analogy in modeling conceptual change

Friedman's (2012) TIMBER model of conceptual change uses structure-mapping over qualitative representations to explain several phenomena in conceptual change. This includes learning intuitive models of force from a series of sketches, where a Companion

ends up going through a sequence of models similar to the sequence of models human students go through (Friedman & Forbus, 2010). The same model has been used to simulate protocol data of children learning about how the seasons work (Friedman, Forbus, & Sherin, 2011), to explain transitions in mental models due to self-explanation in learning about circulatory systems (Friedman & Forbus, 2011), and to explain how feedback might be used to correct misconceptions about the day/night cycle from misunderstood instructional analogies (Friedman, Barbella, & Forbus, 2012).

3.3. Analogical models of cognition are powerful

Stepping back, these models share some important properties. First, they successfully explain or predict a variety of psychological phenomena, including both perceptual and higher order cognitive phenomena. Second, they are effective as machine learning systems. We believe these two points are not unrelated: Building systems that work in a more human like way can provide better learning systems.

While far from complete, we believe that even this suite of examples[9] provides a daunting challenge for those proposing that distributed representations can suffice for explaining human cognition.

4. Why feature vectors are inadequate

Arguments and evidence against spatial models have been known in cognitive psychology for decades (e.g., Tversky 1977), although this point have been ignored in much recent work in AI and cognitive science. This section provides three additional arguments against the sufficiency of feature vectors, and distributed representations more generally, for capturing human cognition. We discuss each in turn.

4.1. Feature vectors do not scale well

Suppose we were to try to use feature vectors to capture the kinds of representations that people use. One impressive feature of human cognition is that it is flexible, combining information from multiple modalities and from far-ranging domains of knowledge. Historical analogies of Donald Trump to Mussolini or to Berlusconi, for example, involve bringing together a wide range of information about economics, personality, and social conditions. To do comparisons such as these, cases must be constructed consisting of many relations, and must be compared. How big are cases? Some statistics on the comparisons involved in five analogy experiments (Forbus et al., 2016) are shown in Table 1.

Given the state of the art in distributed representations, we see two problems. First, there is representing the wide range of relations needed to capture human cognition. In the subset of the Cyc knowledge base[10] we are using, which we expect is smaller than what will be needed to fully capture human cognition, there are over 118,000 attributes,

Table 1
Statistics on representations in five analogy experiments

Task	Mean No. of Entities	Mean No. of Relations
Geometric analogies	2.5	16
Oddity tasks	3	20
Thermodynamics problems	15	89
Physics problems	32	87
Moral reasoning	16	31

23,550 binary relations, and 4,786 ternary relations. A straightforward calculation of representing inputs in terms of this ontology directly would lead to feature vectors in the hundreds of millions to billions of features. Distributed representation schemes provide information compression, to be sure, but to date, none have come close to handling this size of relational vocabulary. Progress is being made, for example Yang, Yih, He, Gao, and Deng (2014, 2015), but they only work with a relatively small number of relations, at least an order of magnitude fewer than even the number of binary relations in ResearchCyc, let alone the encoding of non atomic terms, higher order predicates, and modal operators.

The second scaling problem is that techniques for handling nested structure in distributed representations rely on sequential scanning (Socher et al., 2012; Tai et al., 2015). It is hard to see how such operations can keep up with the speed of human analogical processing, where single-sentence metaphor processing occurs within 1.5 s (Wolff & Gentner, 2011).

The final scaling issue with distributed representations is that they tend to not be incremental. That is, when constructed from words or linked data, all of the data from which they are constructed must be available up front—new terms or predicates cannot be added later, without rerunning the learning process from scratch.[11] This is a significant difference from the incremental nature of human learning.

4.2. Distributed representations often do not lead to human-like performance

While deep learning systems have been successful in applications in speech recognition and machine translation, they require massive amounts of data to do so, far more than a person would experience in a lifetime.[12] For applications this does not necessarily matter, but in terms of explaining cognition it obviously does. When these techniques have been applied to vision, interesting instabilities have been found. By starting with random noise, and using the signal of a fully trained deep learning system as feedback for hill climbing, images can be created that, to people, look like random noise but that the deep learning system will identify with high probability as being an image of a specific type of entity, like "cheetah" (Nguyen, Yosinski, & Clune, 2015). Similarly, small distortions in real images, imperceptible to people, can lead to changing the network's classification of what the image

is (Szegedy et al., 2014). This suggests that such networks are not operating in the way that people do on visual stimuli.[13] As Goodfellow, Shlens, and Szegedy (2015) put it:

> The existence of adversarial examples suggests that being able to explain the training data or even being able to correctly label the test data does not imply that our models truly understand the tasks we have asked them to perform. Instead, their linear responses are overly confident at points that do not occur in the data distribution, and these confident predictions are often highly incorrect. This work has shown we can partially correct for this problem by explicitly identifying problematic points and correcting the model at each of these points. However, one may also conclude that the model families we use are intrinsically flawed.

Attempts to model higher order cognition with distributed representations are rare. One such attempt is the DRAMA model of analogical mapping (Eliasmith & Thagard, 2001), which is claimed to be "an existence proof of using distributed representations to model high-level cognitive phenomena" using holographic reduced representations. We note that DRAMA does not produce candidate inferences, which are an essential capability of analogical matching, which means it would be incapable of serving as a matcher in any of the larger-scale analogical models described above. Another example is the work by Rogers and McClelland (2004) on categorization, lexical access, and how human capabilities degrade with dementia. While the dementia results are intriguing, we note that their models remain small, more like what symbolic models were doing in the 1970s, and training the network requires hundreds to thousands of "epochs," each consisting of rerunning every training example presented to the system. That such models could scale to human-sized conceptual vocabularies seems extremely unlikely.[14]

In machine learning, there are few direct comparisons between structured and distributed representations because the tasks that they are applied to usually do not overlap. However, a recent machine learning task, *link plausibility*, provides a way to compare tradeoffs between structured and distributed representations directly. The goal of this task is to learn to tell whether or not a triple, expressed using semantic web linked data conventions, is likely to be true or not. For example, a true triple is <Barack Obama, Nationality, USA> and a false triple is <Barack Obama, Nationality, Kenya>. Several distributed representation approaches have been tested on this task, on two common databases, compressing structured representations into vectors, matrices, or tensors, depending on the approach. Liang and Forbus (2015) explored a different approach, staying with the original relational form of the data, and using SAGE combined with structured logistic regression to learn an analogy-based classifier. Not only did this system achieve state-of-the-art performance, but it also demonstrated two other advantages. First, it was able to produce understandable explanations for its conclusions, which the other systems could not, because their learned representations are uninspectable. Second, it used orders of magnitude fewer examples than the other systems required. We suspect that this advantage comes directly from staying with relational representations: The other systems, in

effect, "pureed" the symbolic representations into distributed representations, requiring much more training data to implicitly capture the structure that was there to begin with.

4.3. What is being learned in distributed representation systems is not obvious

If what is being learned in the intermediate layers in deep learning systems closely corresponded to the layers of representations that psychophysics and neuroscience suggest are used in human visual processing, then it is hard to see how they could so easily be fooled in ways that people cannot, as noted above. Further evidence of this can be found in the careful analysis of image captioning systems of Devlin, Gupta, Girshick, Mitchell, and Zitnick (2015), which shows that nearest neighbor models often beat deep learning systems. How can this be? They observe that the underlying dataset has many similar images, so captions used for one image are equally appropriate for others. Up to 80% of the captions produced by Vinyals, Toshev, Bengio, and Erhan (2015), for example, could be found explicitly in the training set. These findings cast doubt on just how generative these systems actually are.

This problem is not limited to vision. For example, Levy, Remus, Biemann, and Dagan (2015) looked at nine word representation methods and five lexical-inference datasets used in the literature that have been used in experiments claiming that supervised distributional methods could learn lexical inference relations. They found that these claims did not hold up: That is, instead of learning that "X is a Y," these systems are learning that Y is a good hypernym, independent of X. Similarly, Latent Semantic Analysis (LSA, Landauer & Dumais, 1997) was intended to be a model of semantics, in that the dot product of two vectors would constitute an estimate the similarity of the documents the vectors were produced from. LSA does turns out to be a good predictor of word-word associative priming (Landauer, Foltz, & Laham, 1998), but there is evidence that this is not the same thing as similarity (Gentner & Brem, 1999; Gentner & Gunn, 2001).

Several of the researchers pointing out these problems still believe that distributed representations are an important approach, but that there are problems with current methods for evaluating them. Whether distributed representations can be improved to the point where these issues are overcome is an open question at this point. However, given the problems with distributed representations versus the demonstrated capabilities of relational representations, we suggest that symbolic representations should be a central part of efforts to explain human cognition.

5. Our proposal: Symbolic representations and structure-mapping

We summarized above multiple computational experiments involving structure-mapping over symbolic representations that demonstrate its effectiveness in modeling human cognition. Here we step back and describe our big picture, based on what we have learned so far.

What are the functional advantages of structure-mapping that make it so central to human cognition? Our work, both theoretical and with cognitive models of structure-mapping processes, suggest several reasons:

1. The use of relational representations provides sufficient expressive power for human cognition, unlike feature vectors or multidimensional space models. Structure-mapping processes can operate over structured representations to model problem solving (both conceptual and visual), conceptual change, and moral decision-making, as outlined above.
2. Examples can be immediately reused in new situations, even without a complete and correct domain theory or introducing logical variables (Forbus & Gentner, 1997; Gentner & Medina, 1998). That is, even a partial explanation, manifested in a description by one or more higher order relations (e.g., causality, implication) can give rise to candidate inferences to provide conjectures. This supports within-domain transfer (Klenk & Forbus, 2009). Depending on the nature of the higher order relationships, these inferences can be deductive or abductive (Blass & Forbus, 2016; Forbus, 2015).
3. Structure-mapping supports incremental generalization, via the SAGE model (McLure et al., 2015). As noted above, this enables relational abstractions to be learned with many fewer examples than vector-based models.
4. Analogy solves three of the core problems inherent in probabilistic models, namely determining which aspects of complex stimuli go together (Halstead & Forbus, 2005), where priors come from—the frequency of each statement is tracked within generalizations—and where hypotheses come from (Christie & Gentner, 2010). Thus analogical processing provides a natural complement to Bayesian approaches (e.g., Gershman, Horvitz, & Tennenbaum, 2015).

Looking at the large scale models we have constructed using our models of matching, retrieval, and generalization, a common pattern can be found, one that is the basis for the Companion cognitive architecture (Forbus et al., 2009). In computer science, an architectural stack is a layering of systems that depend on each other. Examples include the hardware and software layers within today's computing systems, web services, and representations in the Semantic Web. It is useful to visualize the patterns of usage for these analogical process models as an *analogy stack*, as show in Fig. 1. Analogical matching (via SME) is the most primitive operation. Similarity-based retrieval (via MAC/FAC), which uses SME in the FAC stage, is next. The case libraries that MAC/FAC operates over are constructed and maintained via analogical generalization (via SAGE).[15]

The generalization pools of SAGE can be considered as analogically constructed models of concepts. SAGE can be used for classification of new examples, using MAC/FAC over a larger pool, the union of generalization pools corresponding to candidate concepts, and tracking from which pool the best retrieval came from. Thus, to the extent that classification is viewed as a central operation in cognition—and it certainly seems necessary,

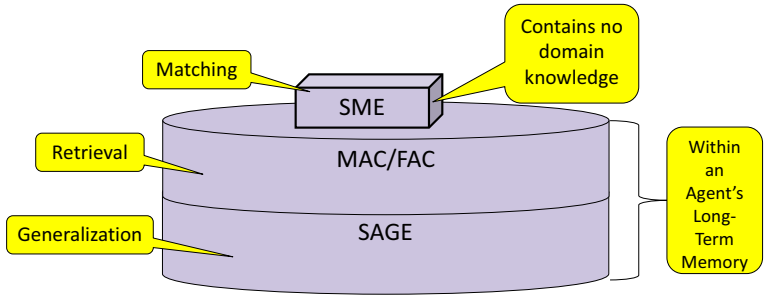

Fig. 1. The analogy stack. SME is used by MAC/FAC, and SAGE uses both SME and MAC/FAC.

but not sufficient—the analogical approach provides a powerful model fully capable of using relational information as well as attributes.

What kinds of concepts might this stack be able to model? Examples we have explored, in addition to those mentioned above, include the following:

1. Learning natural language constructions (McFate & Forbus, 2016; Taylor, Friedman, Forbus, Goldwater, & Gentner, 2011).
2. Word sense disambiguation, where overlap in syntactic and semantic properties of sentences suggest which sense of a word would be most appropriate (Barbella & Forbus, 2013).
3. Decision-making in strategy games, including learning censors (e.g., trying to farm the desert is not a good idea), by accumulating experience as cases and re-applying it (Hinrichs & Forbus, 2007).

In a human-scale analogical cognitive system, the pools might well be divided along functional lines, for example concepts providing word models might have their own copy of MAC/FAC that operates only over that subset of the pool, with visual concepts and concepts used in decision-making similarly operating in parallel. In computer-science terms, most of the computations are *data-parallel*; that is, they can scale arbitrarily, limited only by the size of the underlying substrate. First-principles reasoning would be used for encoding, constraint-checking, planning when prior experience is not available, and some aspects of metacognition. But all of these operations should be facilitated by the rapid availability of relevant generalizations and examples provided via analogical processing.

How many pools and examples might there be? To make a crude lower bound estimate, let us assume the following:

1. A student is exposed to ten new concepts per course, each of which is illustrated by 5 examples and used in five assignments, leading to 10 pools and 100 examples from each course. Suppose the student is on a quarter system, and they take 12 courses per year, across 16 years of schooling. Then schooling would contribute at least 1,920 pools and 19,200 examples. (If education works well, these pools will

overlap pools of prior knowledge, but if not, viz. the inert knowledge problem, they might be distinct.)

2. When doing word sense disambiguation, a generalization pool is used for each pair of word and sense. Using WordNet 3.0 as a baseline,[16] this would suggest at least 360,000 pools, with the number of examples assimilated by them collectively corresponding to the occurrence of each word that a person is exposed to. How many words? Observational studies of families suggest a range from 13 million words to 45 million words in family interactions alone by age 4, depending on socioeconomic status (Hart & Risley, 2003).
3. The number of visual concepts that people are exposed to is even harder to estimate. Using WordNet as a starting point, one ballpark estimate is 1.6 million distinct visual concepts (Smith, 2014). For simplicity, let us assume a mean of 10 examples per visual concept.

These are deliberately conservative estimates—we have not included concepts concerning how to make specific decisions, for example. Roughly, this estimate suggests at least 2 million generalization pools are built up through upward of 45 million examples.

An interesting property of this approach is that it supports a continuum of processing. Recall that examples include both attributes and relations. The attribute component of such descriptions contains the same kind of information as feature vectors do, albeit in a more compact form. Arguably, such descriptions subsume what can be done with feature vectors, since the analogy stack is capable of handling descriptions based purely on attributes. Matching two descriptions with nothing but attributes would still provide a similarity score, like a dot product of feature vectors would. But it would also produce correspondences between the entities and candidate inferences that can be used either to summarize differences or for pattern completion. If these descriptions were accumulated in SAGE, the attribute information would include probabilities. But the same mechanisms also work, and in fact work better, when there is relational information connecting entities in a representation. Moreover, like people, the more overlapping relational information there is, the better it will operate. Thus, this model potentially provides a smooth continuum for reasoning from attribute-based descriptions to rich, relational descriptions laden with causal and inferential information, including statistics that can be used in inferring causality from other information (e.g., Friedman & Forbus, 2008).

5.1. Hybrids: A role for distributed representations?

While we believe there is ample evidence for the necessity of relational representations to explain human cognition, this does not imply that they are sufficient. It could be the case that the combination of analogy and statistics over structured representations is indeed sufficient. But it could also be that some form of distributed representation is used by people in addition to relational representations. For example, Kahneman (2011) among others argues that one way to characterize cognition is in terms of two systems. System 1 is viewed as fast, automatic, emotional, subconscious, and heavily used. System 2 is

viewed as slow, effortful, rule-governed, and conscious. There are many differences between proposed dual-system models, and there are a variety of critiques of them (Evans & Stanovich, 2013). While this distinction might be an oversimplification, it has many attractions, so let us speculate on it. Some confine similarity to System 1, but we suspect that structure-mapping operations are used in both systems. Moreover, one can imagine complementary uses for distributed representations in both systems as well. In System 1, the lack of explanation for many rapid judgments might be due to the use of distributed representations. In System 2, distributed representations might be used as a source of control knowledge used to guide rule-based reasoning. (This seems more of a stretch, since statistics associated with the rules themselves might well suffice [Sharma, Witbrock, & Goolsbey, 2016].)

There is already evidence that such hybrids can be useful in language processing tasks. For example, IBM's Watson includes as one of its evaluation methods the Logical Form Answer Candidate Scorer (LFACS). As described in Murdoch (2011), LFACS is essentially the SME algorithm, but specialized for operating over lexical-level representations connected via syntactic relations. In addition to SME-style structural evaluation, it uses WordNet similarity scores to evaluate competing correspondences. Similarly, Turney (2011) reports on a model using his Latent Relation Mapping Engine (Turney, 2008), which combines structure-mapping and distributed representations to solve word comprehension test problems. Finally, while we have our doubts about word2vec as a model of human word similarity judgments, we used it along with a version of SME applied over syntactic relations computed by the Stanford Core NLP system to achieve state-of-the-art performance on the Microsoft Paraphrase task (Liang, Paritosh, Rajendran, & Forbus, 2016). Similarly, AI2's Aristo system combines statistical methods with reasoning over a semi-automatically constructed knowledge base to answer fourth-grade science test questions (multiple choice only, no diagrams) from unseen, unedited NY Regents Science Exams (Clark et al., 2016). Whether distributed representations turn out to be only a useful engineering approximation or part of human cognition remains an interesting question worth exploring.

6. Conclusions

The current popularity of distributed representations and deep learning is understandable, since they can produce useful results on commercially important problems. However, those representations and processes are unlikely to be sufficient to explain human cognition. They require vastly more data than people need in order to learn, they are unlikely to scale well to the full range of human cognition, their performance can sometimes be a poor fit to what we know about people, and what they are really learning is hard to pin down. In contrast, the combination of analogy and symbolic, relational representations has shown itself to be capable of explaining a wide range of cognitive phenomena, including higher order cognitive capabilities that are well beyond today's deep learning systems.

It is possible to take extreme views. For example, in a recent talk,[17] Geoff Hinton argued that symbols are "the luminiferous aether of cognitive science," which would be replaced by "thought vectors." We might counter with a different historical analogy: Distributed representations are the epicycles of cognitive science. Just as epicycles were the arcane mathematics used to preserve the Earth-centered view of the solar system, distributed representations are ways of preserving an oversimplified model of cognition.

Both analogies might be too harsh. We think the preponderance of evidence available at this point supports the hypothesis that symbolic, relational representations are essential for human cognition. Perhaps hybrid systems may ultimately provide the best account, where distributed representations are not just an engineering approximation. Or it may be that statistics combined with structural representations are sufficient, and distributed representations are wholly unnecessary. Only time will tell.

We end, fittingly, with an analogy: SME is to relational representations as dot product is to feature vectors. We believe this analogy is very important for both cognitive science and for applications of AI and machine learning. The use of structured logistic regression with analogical generalization (Liang & Forbus, 2015) is just one example of exploiting this analogy. We conjecture that many machine learning methods could be translated into analogous analogy-based methods operating over relational knowledge. Combining structural and statistical learning within this framework could provide a very powerful source of new ideas.

Acknowledgments

This work was supported by the Office of Naval Research, through the Intelligent and Autonomous Systems Program and Socio-Cognitive Architectures Program, as well as by the NSF-Funded Spatial Intelligence and Learning Center (SBE-1041707) and the Air Force Office of Scientific Research (FA2386-10-1-4128). We thank Dedre Gentner, Bryan Pardo, Doug Downey, Michael Witbrock, Peter Norvig, Praveen Paritosh, and Johan de Kleer for useful discussions.

Notes

1. See Getoor and Taskar (2007) for an introduction.
2. See Muggleton and de Raedt (1994) for a classic introduction, and the annual ILP conferences (e.g., http://ilp16.doc.ic.ac.uk/) for examples of the latest research.
3. Assuming one game completed per hour, and 12 hours/day of play every day, it would take roughly 6,868 years for a person to play that many games. Clearly, people learn with fewer examples.
4. Hyper-parameters are the parameters of the learning algorithm, rather than the parameters of the learned model.
5. For an engineering perspective on these problems, see Sculley et al. (2015).

6. Digital ink is made up of points, with each stroke represented by a list of points as well as other properties, such as thickness and color.
7. All of the examples from Evans's classic 1968 paper are included in the CogSketch distribution, along with the model.
8. By contrast, Rasmussen and Eliasmith (2011) presented a spiking neuron model that relied on the user to hand-code both the symbolic image representations and the object correspondences between images. No results on Ravens problems were presented for the model.
9. To facilitate experimentation, we have released a corpus of over 5,000 examples of SME comparisons, from a subset of these experiments (Forbus et al., 2016).
10. We are using contents from ResearchCyc 4.0, http://www.cyc.com/platform/researchcyc/. We do not include all of their KB contents, and we have added our own to support qualitative reasoning and analogical reasoning and learning.
11. This is different from pre-training, where, for example, coefficients from a system like word2vec are used as initialization for a new system.
12. Goodfellow, Bengio, and Courville (2016) note, "As of 2016, a rough rule of thumb is that a supervised deep learning algorithm will generally achieve acceptable performance with around 5,000 labeled examples per category, and will match or exceed human performance when trained with a dataset containing at least 10 million labeled examples."
13. See Lee, Ekanadham, and Ng (2008), and Lee, Grosse, Ranganath, and Ng (2009) for attempts to learn more human-like early representations.
14. The set of challenges that the phenomena of analogy raises for distributed representations was laid out in (Gentner and Markman (1993); see also Table 1 of Markman and Gentner (2000). We note that models based on distributed representations have yet to surmount these challenges.
15. For computer science readers: Since we are assuming a vast amount of knowledge and experience in long-term memory, we have flipped the usual way of drawing such a stack; that is, MAC/FAC depends on SME, and SAGE depends on both MAC/FAC and SME.
16. http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html, retrieved 4/23/16.
17. https://drive.google.com/file/d/0B_hicYJxvbiOUHNUUUx6Vl9HRlU/view?pref=2&pli=1, retrieved 4/24/16.

References

Allen, N., & Barnes, B. (2005). Signaling between glia & neurons: Focus on synaptic plasticity. *Current Topics in Neurobiology*, *15*, 542–548.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford, UK: Oxford University Press.

Auld, D., & Robitaille, R. (2003). Glial cells and neurotransmission: An inclusive view of synaptic function. *Neuron*, *40*, 389–400.

Baker, S. (2011) *Final Jeopardy: The story of Watson, the computer that will transform our world*. New York: Houghton Mifflin Harcourt.

Barbella, D., & Forbus, K. (2013). Analogical word sense disambiguation. *Advances in Cognitive Systems*, *2*, 297–315.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*, 1–16.

Barsalou, L. (1999). Perceptual symbol systems. *Behavior and Brain Sciences*, *22*(4), 577–660.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. arXiv:1206.5533v2

Bengio, Y., Lee, D., Bornschein, J., & Lin, Z. (2015) Towards biologically plausible deep learning. arXiv:1502.04156v2.

Blass, J., & Forbus, K.D. (2015). Moral decision-making by analogy: Generalizations vs. exemplars. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX.

Blass, J., & Forbus, K. (2016). Modeling commonsense reasoning via analogical chaining: A preliminary report. Proceedings of CogSci 2016. Philadelphia, PA.

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

Cassimatis, N., Bello, P., & Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher-order cognition. *Cognitive Science*, *32*, 1304–1322.

Chang, M. D., Wetzel, J. W., & Forbus, K.D. (2014). Spatial reasoning in comparative analyses of physics diagrams. In C. Freksa, B. Nebel, M. Hegarty & T. Barkowsky (Eds.), *Spatial Cognition 2014*, LNAI 8684 (pp. 268–282). Cham, Switzerland: Springer International.

Chatterjee, A. (2010). *Disembodying cognition. Language and Cognition*, *2*(1), 79–116.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations via structural alignment. *Journal of Cognition and Development*, *11*(3), 356–373.

Christie, S., Gentner, D., Call, J., & Haun, D. B. M. (2016). Sensitivity to relational similarity and object similarity in apes and children. *Current Biology*, *26*(4), 531–535.

Clark, P., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P., & Khashabi, D. (2016). Combining retrieval, statistics, and inference to answer elementary science questions. Proceedings of AAAI 2016. Phoenix, AZ: AAAI Press.

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

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Collins, A., & Quillian, R. (1970). Does category size affect categorization time? *Journal of Verbal Learning and Verbal Behavior*, *9*(4), 432–438.

Day, S., & Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory and Cognition*, *35*, 39–49.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science*, *311*, 381–384.

Dehghani, M., Tomai, E., Forbus, K., & Klenk, M. (2008). An integrated reasoning approach to moral decision-making. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. Chicago, IL.

DeJong, G. (1993). *Investigating explanation-based learning*. Boston: Kluwer.

Devlin, J., Gupta, S., Girshick, R., Mitchell, M., & Zitnick, L. (2015) Exploring nearest neighbor approaches for image captioning. arXiv:1505.04467v1.

Dietrich, E., & Markman, B. (Eds.) (2000) *Cognitive dynamics: Conceptual and representational change in humans and machines*. New York: Psychology Press.

Domingos, P., Kok, S., Poon, H., Richardson, M., & Singla, P. (2006). Unifying logical and statistical AI. Proceedings of the Twenty-First National Conference on Artificial Intelligence. Boston, MA: AAAI Press.

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

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: a distributed model of analogical mapping. *Cognitive Science*, *25*(2), 245–286.

Eroglu, C., & Barres, B. (2010). Regulation of synaptic connectivity by glia. *Nature*, *468*, 223–231.

Evans, T. (1968). A program for the solution of a class of geometric-analogy intelligence-test questions. In M. Minsky (Ed.), *Semantic information processing* (pp. 223–231). Cambridge, MA: MIT Press.

Evans, J., & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

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

Fan, J., Kalyanpur, A., Gondek, D., & Ferrucci, D. (2012). Automatic knowledge extraction from documents. *IBM Journal of Research and Development*, *56*(3/4), 290–299.

Ferrucci, D. (2012). Introduction to "This is Watson." *IBM Journal of Research and Development*, *56*(3/4), 1–15.

Forbus, K. (1980). Spatial and qualitative aspects of reasoning about motion. Proceedings of the First National Conference on Artificial Intelligence (AAAI-'80), August. Stanford, CA.

Forbus, K. (2001). Exploring analogy in the large. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 23–58). Cambridge, MA: MIT Press.

Forbus, K. (2015). Analogical abduction and prediction: their impact on deception. Proceedings of the AAAI Fall Symposium on Deceptive and Counter-Deceptive Machines. Arlington, VA: AAAI Press.

Forbus, K., Ferguson, R., Lovett, A., & Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*. https://doi.org/10.1111/cogs.12377. [Epub ahead of print]

Forbus, K., & Gentner, D. (1997). Qualitative mental models: Simulations or memories? *Proceedings of the Eleventh International Workshop on Qualitative Reasoning*. Cortona, Italy

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

Forbus, K., Klenk, M., & Hinrichs, T., (2009). Companion cognitive systems: Design goals and lessons learned so far. *IEEE Intelligent Systems*, *24*, 36–46.

Forbus, K., Nielsen, P., & Faltings, B. (1991). Qualitative spatial reasoning: The CLOCK project. *Artificial Intelligence*, *51*, 417–471.

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2011). CogSketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, *3*(4), 648–666.

Friedman, S. E. (2012). Computational conceptual change: An explanation-based approach. Doctoral dissertation. Evanston, IL: Northwestern University, Department of Electrical Engineering and Computer Science.

Friedman, S., Barbella, D., & Forbus, K. (2012). Revising domain knowledge with cross-domain analogy. *Advances in Cognitive Systems*, *2*, 13–24.

Friedman, S., & Forbus, K. (2008). Learning causal models via progressive alignment and qualitative modeling: a simulation. Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci). Washington, D.C.

Friedman, S. E., & Forbus, K. (2010). An integrated systems approach to explanation-based conceptual change. In *Proceedings of the twenty-fourth AAAI conference on artificial intelligence*. Atlanta, GA.

Friedman, S., & Forbus, K. (2011). Repairing incorrect knowledge with model formulation and metareasoning. Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain.

Friedman, S. E., Forbus, K. D., & Sherin, B. (2011). Constructing and revising commonsense science explanations: A metareasoning approach. Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems. Arlington, VA: AAAI Press.

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

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. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34* (5), 752–775.

Gentner, D., & Brem, S. (1999) Is snow really like a shovel? Distinguishing similarity from thematic relatedness. In M. Hahn & S.C. Stoness (Eds.), *Proceedings of the Twenty-first Annual Meeting of the cognitive Science Society* (pp. 179–184). Mahwah, NJ: Lawrence Erlbaum Associates.

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory and Cognition*, *29*(4), 565–577.

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* (pp. 855–862). San Mateo, CA: Kaufmann.

Gentner, D., & Markman, A. B. (1995). Similarity is like analogy: Structural alignment in comparison. In C. Cacciari (Ed.), *Similarity in language, thought and perception* (pp. 111–147). Brussels: BREPOLS.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*, 263–297.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*(6), 297–301.

Gershman, S., Horvitz, E., & Tennenbaum, J. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Getoor, L., & Taskar, B. (2007). *An introduction to statistical relational learning*. Cambridge, MA: MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Goodfellow, I., Shlens, J, & Szegedy, C. (2015) Explaining and harnessing adversarial examples. Proceedings of ICLR 2015. San Diego, CA.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. Arxiv:1410.5401v2.

Halstead, D., & Forbus, K. (2005). Transforming between propositions and features: bridging the gap. Proceedings of the 20th AAAI Conference on Artificial Intelligence. Pittsburgh, PA: AAAI Press.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by Age 3. *American Educator*, 4–9.

Hinrichs, T., & Forbus, K. (2007). Analogical learning in a turn-based strategy game. In M. Veloso (Ed.), *Proceedings of the twentieth international joint conference on artificial intelligence* (pp. 853–858). Hyderabad, India: International Joint Conferences on Artificial Intelligence.

Hinton, G. (1979). Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, *3*(3), 231–250.

Hofstadter, D., & Sander, E. (2013). *Surfaces and essences: Analogy as the fuel and fire of thinking*. New York: Basic Books.

Holden, M. P., Newcombe, N. S., & Shipley, T. F. (2015). Categorical biases in spatial memory: The role of certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 473–481.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: The MIT Press.

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*, 109–118.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Huttenlocher, J., Hedges, L., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*, 352–376.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Klenk, M., & Forbus, K. (2007). Cognitive modeling of analogy events in physics problem solving from examples. Proceedings of CogSci-07. Nashville, TN.

Klenk, M., & Forbus, K. (2009). Analogical model formulation for AP physics problems. *Artificial Intelligence*, *173*, 1615–1638.

Klenk, M., Forbus, K., Tomai, E., & Kim, H. (2011). Using analogical model formulation with sketches to solve Bennett Mechanical Comprehension Test problems. *Journal of Experimental and Theoretical Artificial Intelligence*, *23*(3), 299–327.

Kokinov, B., & French, R. M. (2003). Computational models of analogy-making. In L. Nadel (Ed.), *Encyclopedia of cognitive science*. Vol. *1* (pp. 113–118). London: Nature Publishing Group.

Krawczyk, D., Holyoak, K., & Hummel, J. (2005). The one-to-one constraint in analogical mapping and inference. *Cognitive Science*, *29*, 797–806.

Kuehne, S., Gentner, D., & Forbus, K. (2000). Modeling infant learning via symbolic structural alignment. Proceedings of the Cognitive Science Society. Philadelphia, PA.

Kuipers, B. (2000). The Spatial Semantic Hierarchy. *Artificial Intelligence*, *119*(191–233), 2000.

Laird, J. (2012). *The SOAR cognitive architecture*. Cambridge, MA: MIT Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *512*, 436–444.

Lee, H., Ekanadham, C., & Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. In D. Lenat & R. Guha (Eds.), Advances in neural information processing systems (pp. 873–880). *Building large knowledge-based systems*. Boston: Addison-Wesley.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In L. Bottou & M. Littman (Eds.), *Proceedings of the 26th annual international conference on machine learning* (pp. 609–616). New York: ACM.

Levy, O., Remus, S., Biemann, C., & Dagan, I. (2015) Do supervised distributional methods really learn lexical inference relations? Proceedings of NAACL 2015. Denver, CO.

Liang, C., & Forbus, K. (2015). Learning plausible inferences from semantic web knowledge by combining analogical generalization with structured logistic regression. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas.

Liang, C., Paritosh, P., Rajendran, V., & Forbus, K. (2016) Learning paraphrase identification with structural alignment, Proceedings of IJCAI-2016. New York, NY.

Lockwood, K., Lovett, A., & Forbus, K. (2008). Automatic classification of containment and support spatial relations in English and Dutch. In C. Freksa, N. Newcombe, P. Gardenfors & S. Wolfl (Eds.), Proceedings of the International Conference on Spatial Cognition VI: Learning, Reasoning, and Talking about Space (pp. 283–294). Freiburg, Germany: Springer.

Lovett, A. (2012). Spatial routines for sketches: A framework for modeling spatial problem-solving. Doctoral dissertation. Evanston, IL: Northwestern University, Department of Electrical Engineering and Computer Science.

Lovett, A., & Forbus, K. (2011). Cultural commonalities and differences in spatial problem solving: A computational analysis. *Cognition*, *121*, 281–287.

Lovett, A., & Forbus, K. (2012). Modeling multiple strategies for solving geometric analogy problems. Proceedings of the 34th Annual Conference of the Cognitive Science Society. Sapporo, Japan.

Lovett, A., & Forbus, K. (2013). Modeling spatial ability in mental rotation and paper-folding. Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. 930–935). Berlin, Germany.

Lovett, A., & Forbus, K. (2017). Modeling visual problem-solving as analogical reasoning. *Psychological Review*, *124*(1), 60. https://doi.org/10.1037/rev0000039.

Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. Proceedings of CogSci-10. Portland, OR.

Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive Science*, *33*(7), 1192–1231.

Marcus, G., Vijayan, S., Bandi Rao, S., & Vishton, P. (1999). Rule-learning in seven-month-old infants. *Science*, *283*, 77–80.

Markman, A. (1997). Constraints on analogical inference. *Cognitive Science*, *21*, 373–418.

Markman, A. (1998). *Knowledge Representation*. Mahwah, NJ: Lawrence Erlbaum.

Markman, A., & Gentner, D. (1993a). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467.

Markman, A. B., & Gentner, D. (1993b). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*, 517–535.

Markman, A., & Gentner, D. (2000). Structure-mapping in the comparison process. *American Journal of Psychology*, *113*(4), 501–538.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

McCord, M., Murdock, J., & Boguraev, B. (2012). Deep parsing in Watson. *IBM Journal of Research and Development*, *56*(3/4), 264–278.

McFate, C., & Forbus, K. (2016). Analogical generalization and retrieval for denominal verb interpretation. Proceedings of CogSci 2016. Philadelphia, PA.

McLure, M.D., Friedman, S.E., & Forbus, K.D. (2015). Extending analogical generalization with near-misses. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX.

Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2007). BLOG: Probabilistic models with unknown objects. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.

Minton, S., Carbonell, J., Knoblock, C., Kuokka, D., Etzioni, O., & Gil, Y. (1989). Explanation-based learning: A problem-solving perspective. *Artificial Intelligence*, *40*(1–3), 63–118.

Mitchell, T., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, *1*, 47–80.

Muggleton, S., & de Raedt, L. (1994). Inductive logic programming: theory and methods. *Journal of Logic Programming*, *19–20*(1), 629–679.

Murdoch, J. (2011). Structure-mapping for jeopardy! clues. In A. Ram & N. Wiratunga (Eds.), *ICCBR 2011*, LNAI 6880 (pp. 6–10). Berlin: Springer-Verlag.

Nguyen, A., Yosinski, J., & Clune, J. (2015) Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015, Boston, MA.

Ouyang, T., & Forbus, K. (2006). Strategy variations in analogical problem solving. Proceedings of the 21st AAAI Conference on Artificial Intelligence, Boston, MA.

Palmer, S. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and non-human minds. *Behavioral and Brain Sciences*, *31*, 109–178.

Perrott, D. A., Gentner, D., & Bodenhausen, G. V. (2005). Resistance is futile: The unwitting insertion of analogical inferences in memory. *Psychonomic Bulletin & Review*, *12*, 696–702.

Rasmussen, D., & Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science*, *3*(1), 140–153.

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

Rosenbloom, P. S. (2010). An architectural approach to statistical relational AI. Proceedings of the AAAI-10 Workshop on Statistical Relational AI. Atlanta: AAAI Press.

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36* (6), 1019–1050.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 2494–2502.

Sharma, A., Witbrock, M., & Goolsbey, K. (2016) Controlling search in very large commonsense knowledge bases: a machine learning approach. Proceedings of the Fourth Conference on Advances in Cognitive Systems. Evanston, IL.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489.

Smith, J. R. (2014). How many visual concepts?. *IEEE Multimedia*, *21*(1), 2–3.

Socher, R., Huval, B., Manning, C., & Ng, A. (2012). Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*. arxiv:1312.6199.

Tai, K.S., Socher, R., & Manning, C.D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China.

Taylor, J. L. M., Friedman, S. E., Forbus, K. D., Goldwater, M., & Gentner, D. (2011). Modeling structural priming in sentence production via analogical processes. Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci). Boston, MA.

Tomai, E., & Forbus, K. (2009). EA NLU: Practical language understanding for cognitive modeling. Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference. Sanibel Island, Florida

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Turney, P. (2008). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, *33*, 615–655.

Turney, P. (2011). Analogy perception applied to seven tests of word comprehension. *Journal of Experimental & Theoretical Artificial Intelligence*, *23*(3), 343–362.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

VanLehn, K., & Jones, R. (1991). Learning physics via explanation-based learning of correctness and analogical search control. In L. Birnbaum, & G. Collins (Eds.), *Machine learning: Proceedings of the 8th international workshop* (pp. 110–114). San Mateo, CA: Morgan-Kaufmann.

Veloso, M., Carbonell, J., Perez, A., Borrajo, D., Fink, E., & Blythe, J. (1995). Integrated planning and learning: The PRODIGY architecture. *Journal of Theoretical and Experimental Artificial Intelligence*, *7*, 81–120.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: a neural image caption generator. Proceedings of CVPR 2015, Boston, MA.

Wolff, P., & Gentner, D. (2011). Structure-mapping in metaphor comprehension. *Cognitive Science*, *35*, 1456–1488.

Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2014) Learning multi-relational semantics using neural-embedding models. NIPS 2014 Workshop on Learning Semantics. Montreal, Canada.

Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2015) Embedding entities and relations for learning and inference in knowledge bases. International Conference on Learning Representations (ICLR). San Diego, CA.