COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 40 (2016) 1128–1162

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12265

Concepts as Semantic Pointers: A Framework and Computational Model

Peter Blouw, Eugene Solodkin, Paul Thagard, Chris Eliasmith

Center for Theoretical Neuroscience, University of Waterloo

Received 9 July 2013; received in revised form 24 February 2015; accepted 20 March 2015

Abstract

The reconciliation of theories of concepts based on prototypes, exemplars, and theory-like structures is a longstanding problem in cognitive science. In response to this problem, researchers have recently tended to adopt either hybrid theories that combine various kinds of representational structure, or eliminative theories that replace concepts with a more finely grained taxonomy of mental representations. In this paper, we describe an alternative approach involving a single class of mental representations called "semantic pointers." Semantic pointers are symbol-like representations that result from the compression and recursive binding of perceptual, lexical, and motor representations, effectively integrating traditional connectionist and symbolic approaches. We present a computational model using semantic pointers that replicates experimental data from categorization studies involving each prior paradigm. We argue that a framework involving semantic pointers can provide a unified account of conceptual phenomena, and we compare our framework to existing alternatives in accounting for the scope, content, recursive combination, and neural implementation of concepts.

Keywords: Concepts; Categorization; Neural computation; Semantics; Computational modeling; Mental representation

1. Introduction

The study of concepts has played a central role in the advancement of recent theories of cognitive function. Phenomena ranging from categorization to language use have been profitably described in terms of conceptual processing (see Murphy, 2002, for a review), and many influential descriptions of cognitive development have been produced on the assumption that concepts are the basic representational entities that comprise our

Correspondence should be sent to Peter Blouw, Center for Theoretical Neuroscience, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1. E-mail: pblouw@uwaterloo.ca

knowledge of the world (e.g., Carey, 1985, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). However, despite the obvious importance of concepts to ongoing work in the cognitive sciences, researchers have developed very different views regarding how concepts are structured and represented in the brain.

A primary reason for this disagreement is that theorists working in different disciplines have numerous and sometimes conflicting explanatory goals. Psychologists, for example, typically wish to explain empirical data from experiments involving tasks like categorization and concept learning (e.g., Lin & Murphy, 1997; Regehr & Brooks, 1993; Rips, 1989; Smith & Medin, 1981). Philosophers, on the other hand, typically wish to understand the semantics and possession conditions of concepts (e.g., Fodor, 1998; Laurence & Margolis, 1999; Peacocke, 1992; Prinz, 2002). Even when a set of explanatory goals is agreed upon, the scope of the data to be accounted for is often too vast and disparate to allow for the development of a unified theory (Murphy, 2002). Categorization phenomena alone, for example, cannot be comprehensively explained using the individual resources of prototype theories, exemplar theories, or theory theories of concepts (Rogers & McClelland, 2004).

To deal with this impasse, researchers have most recently tended to adopt one of two general strategies. The first strategy involves the proposal of "mixed" or "pluralistic" models in which individual concepts correspond to a number of related or co-referring representational structures that account for different phenomena (Laurence & Margolis, 1999; Murphy, 2002; Weiskopf, 2009). The second strategy, in contrast, involves eliminating the term "concept" from the vocabulary of the cognitive sciences in favor of a taxonomy of more finely grained mental representations that each serve distinct functions (Machery, 2009). The relative merits of these two approaches are the subject of ongoing debate (Machery, 2010), but it is safe to say that neither view has achieved widespread support.

In this paper, we propose an alternative, unifying solution to the current challenges in concept research. Using methods for characterizing representational states in neural systems (Eliasmith, 2003, 2013; Eliasmith & Anderson, 2003), we describe concepts in terms of processes involving a recently postulated class of mental representations called "semantic pointers" (Eliasmith, 2013). Roughly speaking, semantic pointers are neurally instantiated, symbol-like representations that can be transformed in numerous ways to yield further representations that function to support cognitive processes like categorization, inference, and language use. Notably, semantic pointers have been successfully used to account for a range of perceptual, cognitive, and motor behaviors in what is currently the world's largest functional model of the human brain (Eliasmith et al., 2012). However, this past work does not explore the relevance of semantic pointers to conceptual phenomena in detail. Our aim here, accordingly, is to show that a modeling framework based on semantic pointers can offer a unified explanation of the kinds of phenomena that concept theorists have traditionally been interested in. To support our claims, we describe a biologically plausible spiking neuron model that processes semantic pointers to account for data from categorization experiments that have been used to bolster three competing accounts of concepts: prototype theory, exemplar theory, and theory theory.

2. Criteria for a theory of concepts

Although the criteria by which theories of concepts are evaluated are often controversial, it is widely acknowledged that certain cognitive functions are paradigmatically *conceptual* functions. For example, language use, inference, and the formation of propositional attitudes are just a few of many cognitive tasks uncontroversially defined in terms of operations involving concepts. We accordingly take as our starting point the idea that a theory of concepts ought to explain how these conceptual functions are implemented. However, in addition to these functional explanations, certain theoretical explanations should also be provided. For example, one ought to explain how concepts can range in kind from the abstract to the ordinary, and explain how they can refer to groups of objects in the world. With these considerations in mind, we propose the following criteria as minimal requirements for a satisfactory account of conceptual processing (cf. Barsalou, 1999; Fodor, 1998; Laurence & Margolis, 1999; Prinz, 2002):

1. Categorization
2. Recursive binding
3. Neural implementation
4. Scope
5. Content

There are, of course, other criteria one might choose, but we have selected these five for the simple reason that they seem to capture properties common to a large number of conceptual phenomena. Categorization tasks, for instance, are widely studied in the literature on concepts (Murphy, 2002), and in some cases recruit background knowledge in the form of inferences that relate object properties to category membership. Thus, it is plausible that descriptions of additional conceptual processes involving inference and language can be understood partly as more complex forms of categorization. We focus exclusively on categorization effects in our simulations because of these and related considerations.[1]

An explanation of binding, too, is an important goal for any account of conceptual processing: it underlies both the formation of compositional structures involving multiple concepts (e.g., LARGE RED DOG) and the integration of multimodal representations of category instances. Binding has attracted a great deal of attention amongst researchers interested in the structure of mental representations (e.g., Jackendoff, 2002), and we thus take it to be a somewhat uncontroversial constraint.

Regarding neural implementation, it is of course something of a platitude to say that conceptual processes are neural processes. But since the nature of neural processes likely constrains the types of functions that can be easily computed by a cognitive system (Eliasmith & Anderson, 2003), it remains an open question whether the functions described by any particular cognitive model are, in fact, neurally implementable. So, the adoption of a neural implementation criterion suggests that, all else being equal, a demonstration of the implementation of a particular model counts considerably in its favor.

As for the more theoretical criteria, scope refers to the broad range of different kinds of concepts. There are concepts for perceivable objects (e.g., TABLE), abstractions (e.g., VIRTUE), theoretical posits (e.g., GENE), mathematical terms (e.g., SUM), and non-existent entities (e.g., CENTAUR), among other things (Prinz, 2002). A good theory should be able account for these different classes of concepts, and it should also be consistent with available evidence implicating particular neural systems and anatomical regions with the processing of these classes. For instance, studies of neurological patients with semantic deficits suggest that concepts for concrete and abstract entities are processed in distinct neural systems (Shallice & Cooper, 2013).

Finally, the fact that concepts are about things means that they have content or meaning. This content, in turn, can be roughly defined in terms of how a given concept characterizes what it represents. An adequate theory must explain why a given concept refers to some things and not others (i.e., provide an account of its extension), and it must also explain why this concept describes these referents in some ways and not others (i.e., provide an account of its intension). The philosophical literature on the semantics and individuation of mental representations provides the motivation for adopting this criterion (e.g., Fodor, 1987, 1998; Peacocke, 1992; Prinz, 2002).

In summary, the first three criteria concern the nature and implementation of conceptual functions, while the last two criteria concern theoretical properties of the representations that enable these functions. Our framework is designed to meet these criteria, although our discussion of the first three is meant to be more comprehensive, while our discussion of the last two is meant to be more suggestive. To begin developing the framework in detail, we first describe the principles of neural representation and computation that motivate a number of our arguments.

3. Neural representation and computation

While it is widely accepted that mental representations are features of neural systems, current approaches to cognitive modeling do not generally characterize representations in highly detailed neural terms. Symbolic approaches, for instance, typically describe mental phenomena in terms of computations defined over atomic representations structured by a language-like syntax (e.g., Fodor, 1975); neural details are rarely, if ever, a consideration. Connectionist approaches, in contrast, characterize representations in sub-symbolic terms using weighted connections between large numbers of individual processing nodes (e.g., Rogers & McClelland, 2004; Rumelhart & McClelland, 1986). These models, however, only roughly correspond to the structure of the brain, and they leave out a number of important details regarding the physiological properties, dynamical properties, and connectivity of real neurons.

We favor an approach that describes representation and computation in terms of the activities of large numbers of individually spiking neurons. More specifically, we adopt the Neural Engineering Framework (NEF) developed by Eliasmith and Anderson (2003). According to this framework, patterns of activity in spiking neurons can be characterized

using mathematical objects such as vectors (i.e., sets of numerical values), which in turn capture information about the world (via the tuning of the neurons to environmental stimuli).[2] By specifying sets of synaptic weights between two or more populations of neurons, transformations of such vectors can be computed, including transformations that bind together multiple vectors to embed complex hierarchical structures in a vector space. It is for this reason that the NEF is occasionally characterized as a compiler that translates algorithms defined over vectors into a language of neural spikes (see Eliasmith, 2013). For our purposes, there are two significant advantages to characterizing the behavior of neural systems in this quantitative manner. First, there are well-established techniques for translating various kinds of lexical, sensory, and motor representations into vectors (Georgopoulos, Schwartz, & Kettner, 1986; Jones & Mewhort, 2007; Plate, 2003). Vectors accordingly have the necessary representational power to account for the multimodal nature of conceptual representations (cf. Barsalou, 1999). Secondly, computations involving vectors can be used to implement powerful forms of recursive binding (Gayler, 1998; Kanerva, 1994; Plate, 2003; Smolensky, 1990). Given the vast array of contents associated with even the simplest concepts, an account of binding is likely essential for the development of a plausible model of conceptual processing.

Binding can be carried out in the NEF using a process called circular convolution (Eliasmith, 2004, 2013). Leaving the mathematical details aside, circular convolution can be thought of as a function that blends two input vectors into a single output vector of the same dimensionality. Implementing this function is relatively straightforward: If two "input" neural populations each representing a vector are connected to an intermediary population that projects to an "output" population, one can use the NEF to solve for a set of synaptic weights between the populations that will result in the output population encoding a vector that is the convolution of the two input vectors. This process can be repeated indefinitely, and it can also be reversed to recover an approximation of any one vector bound into such a recursively generated structure.[3] Overall, the NEF has all of the tools needed to describe highly complex syntactic operations involving the composition and decomposition of a diverse range of neural representations. In combination, the NEF principles of neural representation and computation allow for the description of a very powerful kind of representation that Eliasmith (2013) refers to as a "semantic pointer." We use the notion of a semantic pointer as a starting point in developing an account of concepts that can adequately satisfy all five of the criteria introduced in Section 2.

4. Semantic pointers

In its most basic form, a semantic pointer can be thought of as a compressed representation that captures summary information about a particular domain. Typically, such representations derive from perceptual inputs. An image of an object in one's visual field, for instance, will initially be encoded as a pattern of activity in a very large population of neurons. Through transformations of the sort described above, however, further layers of neural populations produce increasingly abstract statistical summaries of the original

visual input (see Fig. 1). Eventually, a highly compressed representation of the input can be produced. Such a characterization is consistent both with the decrease in the number of neurons found in later hierarchical layers of the visual cortex, and with the development of neurally inspired hierarchical statistical models for dimensionality reduction (Hinton & Salakhutdinov, 2006; Serre, Oliva, & Poggio, 2007). Analogous representations can be generated in other modalities such as audition and sensation.

The reason compressed representations of this sort are called semantic pointers is because they retain semantic information about the states they represent by virtue of being non-arbitrarily related to these states through the compression process. The reason why the representations are referred to as pointers is because they can be used to "point to" or regenerate representations at lower levels in the compression network (Hinton & Salakhutdinov, 2006). Moreover, any given semantic pointer can be manipulated independently of the network that is used to generate it. A semantic pointer of a table percept, for example, could be used in cognitive tasks related to tables without necessarily prompting a reactivation of the richer perceptual representations at the bottom of the relevant compression network.

The computational power of semantic pointers lies in their ability to be bound together (using compression operations such as circular convolution) into highly structured representations containing lexical, perceptual, and motor information from a variety of sources. Importantly, such structured representations are themselves semantic pointers, because

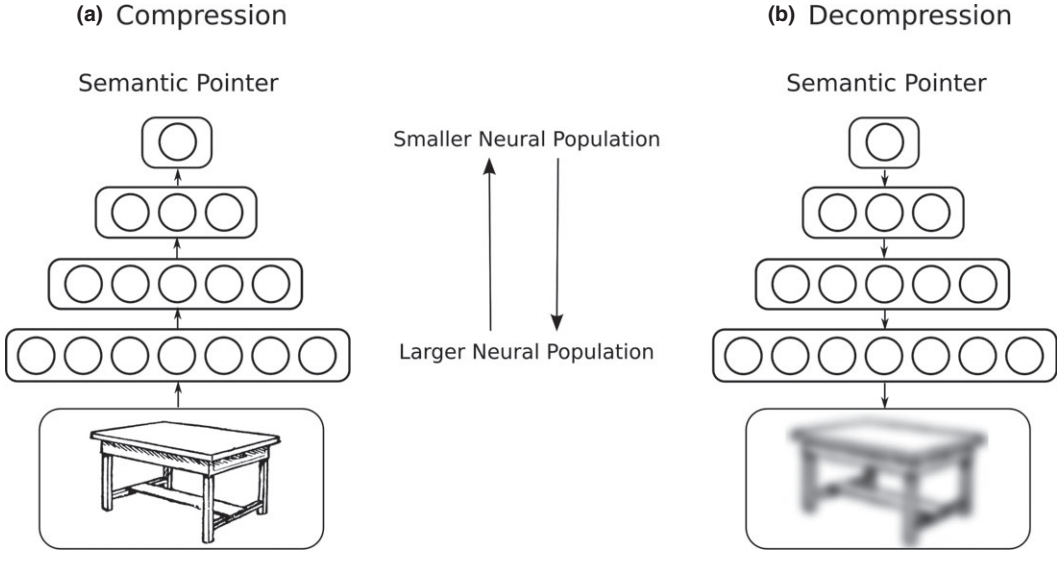

Fig. 1. Hierarchical populations of neurons used for the compression and decompression of perceptual data. The number of nodes in each layer corresponds to the dimensionality of the representation. The low dimensional semantic pointer at the top of the hierarchy in (a) is thus a compressed representation of a percept, and the high dimensional representation at the bottom of the hierarchy in (b) amounts to a partial recovery of this percept from the appropriate semantic pointer.

they can point to and regenerate the subordinate representations from which they are built. Consider again the toy example of a table. Using recursive binding of the sort already described, which is a compression operation, semantic pointers for visual and tactile images of tables could be combined, along with pointers for an auditory image of the sound "table" and a visual image of the letters "t-a-b-l-e." Additionally, various structures corresponding to verbal information like "has a flat surface" or "used for eating meals" might also be bound together. These structures would themselves be built out of other semantic pointers, including compressed visual images of flat surfaces, meal settings, and so on. Overall, the result of these numerous binding operations is a single representation that captures relations among a wide range of table-related contents. And this single representation can be transformed in numerous ways to re-access images of tables, verbal information about tables, or motor commands commonly used to interact with tables.

At this point, it should be clear that semantic pointers are highly applicable to the explanation of conceptual phenomena. They can account for symbolic processes, perceptual simulations, and a host of other functions all centered upon a single object class. In other words, they can act as a summary representation of a category of things in the world, which is precisely what a concept is often taken to be. Successful neural simulations of conceptual tasks such as simple linguistic inference (Eliasmith, 2013), inductive reasoning (Rasmussen & Eliasmith, 2011), and rule-based problem solving (Stewart & Eliasmith, 2011) have all been produced using semantic pointers, along with a large-scale brain model capable of executing a variety of cognitive functions (Eliasmith et al., 2012). Based on these successful applications, we think that the notion of a semantic pointer provides an ideal foundation for accounting for a wide range of conceptual phenomena.

5. Concepts as semantic pointers

It is tempting to claim that concepts just are semantic pointers. However, we avoid this theoretical formulation for a simple reason: Semantic pointers cannot meet all of the desired criteria when considered in isolation. Recall that a semantic pointer is simply a vector encoded by the spiking activity in a population of neurons. This vector captures relations between a wide range of other representations, and it can be transformed in various ways to access these representations, but the vector itself does not possess anywhere near the full semantic content of an ordinary concept. It is better, then, to think of a semantic pointer as an entity that enables the occurrence of a concept rather than as an entity that is equivalent to a concept.

On our account, concepts are best thought of dispositionally. To possess a concept is to be able to activate various sequences of neural states that correspond to things like visual and auditory simulations, expressions of natural language, and motor commands all centered on a single category. The interrelated neural states that feature in these processes result from the transformation of semantic pointers, and on any given occasion in which a particular concept occurs, only a limited range of all possible transformations will be carried out. In other words, the neural processes that comprise the occurrence of a given

concept are context and task dependent (Barsalou, 1999). Conceptual tasks involving verbal reasoning will, for instance, activate different neural states than conceptual tasks involving the categorization of tactile stimuli (see Fig. 2). Likewise, a cognitive task involving the concept DOG might invoke a visual simulation of a large animal in one individual while invoking a simulation of a smaller animal in another individual. Part of the burden of elaborating on this theory is to give an account of the factors that influence such contextual variability. Nonetheless, our claim is that these various neural states that are constitutive of conceptual processing stem from a common point of origin: namely, the transformation of a semantic pointer.

Given this description, it is important to consider the obvious similarities between our view and recently developed "neo-empiricist" accounts that identify conceptual processing with the partial re-activation of previously captured perceptual states (e.g., Barsalou, 1999; Barsalou, Simmons, Barbey, & Wilson, 2003; Barsalou, Santos, Simmons, & Wilson, 2008; Prinz, 2002). Barsalou (1999), for example, identifies concepts with "simulators" or organized systems of category-specific perceptual symbols that can be selectively transferred into working memory. As one might expect, this transfer of symbols into working memory is highly analogous to the transformation of a semantic pointer to access detailed perceptual representations. Similarly, Prinz's (2002) claim that concepts are "proxytypes" or "perceptually derived representations that can be recruited by

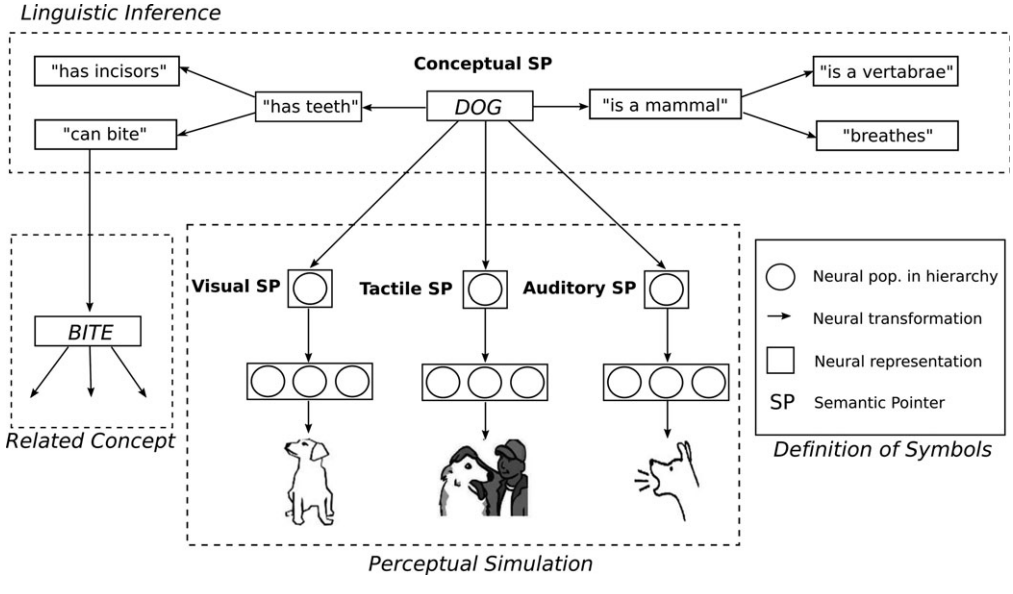

Fig. 2. A simplified diagram of possible transformations of a semantic pointer for the concept DOG. Transforming the semantic pointer can result in perceptual simulations, linguistic inferences, and the consideration of related concepts. An occurrence of the concept DOG, in our account, is a process through which a set of these possible transformations is realized. All the representations and transformations depicted in this diagram are compatible with the principles of neural implementation described in Section 3.

working memory to represent a category" (p. 149) shares much with our notion of concepts as processes corresponding, in part, to perceptual simulations.

There are two significant differences between the semantic pointer framework and these neo-empiricist accounts. For one, the semantic pointer theory is consistent with the existence of amodal representations. A semantic pointer corresponding to a lexical term, for example, can contain amodal statistical information regarding the co-occurrence of related terms in certain contexts (Eliasmith, 2013). Access to such information in the absence of perceptual representations can explain response times during certain tasks that require the verification of matches between words and properties (e.g., Solomon & Barsalou, 2004). Moreover, given the controversial nature of the arguments and evidence commonly cited in support of strictly empiricist accounts (Machery, 2007), and the existence of evidence suggesting that concrete and abstract concepts are processed in at least partially separable systems (Shallice & Cooper, 2013), we see the neutrality of the semantic pointer framework on this issue as a virtue.

Second, the view we propose offers a distinct account of how concepts are actually represented. Neo-empiricist accounts generally identify concepts with either (1) simulators (i.e., organized systems of perceptual symbols) or (2) simulations (i.e., temporary representations in working memory). The problem with the first option is that it is underspecified with respect to how simulators carry out their functions (Dennett & Viger, 1999). The semantic pointer framework addresses this shortcoming by identifying concept occurrences with neural processes and by mechanistically describing these processes in terms of independently motivated principles of neural computation and a model-based implementation, described below. The problem with the second option is that it entails that each instance of a specific simulation corresponds to a unique concept. If so, then one can have multiple concepts that denote a single category, and one would rarely elicit the same concept twice, since the production of identical simulations across time is unlikely. Without some explanation of why such simulations form distinct concepts (or why they are related if they are not in fact distinct concepts), the theory is left vague and imprecise. Our account avoids this problem by unifying diverse occurrences of a single concept through the postulation of a common underlying neural mechanism.

Further similarities are also evident between the semantic pointer framework and connectionist models that are trained through gradient descent to associate various modality-specific representations (e.g., percepts, verbal descriptions, etc.) via mediating, amodal semantic representations (Rogers et al., 2004; Roy & Pentland, 2002; Rumelhart & McClelland, 1986). Two key features differentiate our approach. The first is the account of representational binding that we provide. Connectionist models that learn to associate representations have no means by which to bind representations recursively, and thus no means by which to account for compound concepts or syntactically structured representations. Second, a form of localist encoding is often used wherein individual processing nodes are taken to represent unique linguistic predicates and perceptual features. This encoding violates the implementation criterion, because the nature of the correspondence between these nodes and the neural substrate is left unspecified.

For all of these reasons, we propose that our account offers a promising new approach to satisfying the criteria introduced in Section 2. However, before returning to an evaluation of the theory with respect to these criteria, we first describe a semantic pointer-based model that is able to account for important features of the prototype, exemplar, and theory-theory accounts of concepts. While this computational model is not comprehensive, it provides a good initial demonstration of the potential for a semantic pointer-based framework to offer a unified explanation of conceptual phenomena.

6. Model description

We focus our modeling efforts on three paradigmatic categorization studies. The first study, conducted by Posner and Keele (1968), suggests that when subjects learn to categorize patterns of dots generated through the disturbance of various prototype patterns, they abstract and utilize information about the relevant prototype. The second study, conducted by Regehr and Brooks (1993), demonstrates that similarity-based categorization strategies can override more analytic strategies under certain circumstances. The third study, conducted by Lin and Murphy (1997), investigates the effects of background knowledge on categorization decisions involving identical visual stimuli given distinct functional descriptions. Together, these studies catalogue a diverse range of categorization effects.

We propose a single model architecture that does not change when accounting for any of the different effects. Functionally, the model receives vectors corresponding to compressed natural images as input and produces vectors corresponding to motor responses as output. All intermediate processing is implemented using approximately 300,000 simulated leaky-integrate-and-fire (LIF) neurons, and 128 dimensional vectors are used in every simulation. Details regarding how the LIF neurons are used to encode, decode, and transform vectors can be found in Part A of the Supplemental Materials.

At a more specific level, the model architecture involves a working memory system, an action selection system, and two further subsystems that implement perceptual and inferential evaluations of input stimuli (see Fig. 3). Working memory stores semantic pointers that encode visual exemplars and soft rules that define category membership. The action selection system controls how information is extracted from these semantic pointers and used to manipulate input stimuli in a task-dependent manner. Anatomically, the action selection system is mapped on to portions of basal ganglia and thalamus, while the other subsystems are mapped on to cortex. These anatomical mappings are largely motivated by other work (Eliasmith, 2013; Eliasmith et al., 2012) and for present purposes are best viewed as plausible assumptions. The functionality of the model is our primary concern, and in the simulations reported below, we do not use the model to account for neural data in a way that would independently justify the mappings in question.

During each experiment, the model is presented with a visual input that cues the current experimental task, followed by a stimulus to categorize. Based on the cue, the action selection system initiates processes that decompress a semantic pointer to either

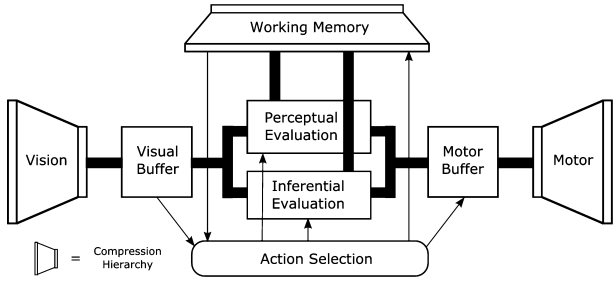

Fig. 3. Functional architecture of the model. Thick black lines indicate connections between subsystems, and thin lines with arrows indicate connections that allow the action selection system to monitor and modify representational states in these subsystems. During each run, the model is presented with a vector indicating the current task, followed by a vector corresponding to a compressed image of a stimulus. The task vector triggers actions that decompress a semantic pointer stored in working memory to obtain perceptual or inferential information that enables categorization of the stimulus. Categorization judgments are routed through the motor buffer as output. The visual and motor compression hierarchies are not modeled in the simulations below (cf. Eliasmith et al., 2012), but the former is taken to compress images into semantic pointers, while the latter is taken to decompress semantic pointers into motor activities. Subsystems contain internal components that perform simple processing and routing tasks as described in the main text.

(a) compare the stimulus to previously learned exemplars or (b) apply a set of rules that define category membership.

To provide more formal details, the working memory system contains neural populations whose activities represent semantic pointers corresponding to concepts learned during the training phase of each experiment. In the case of experiments involving perceptual categorization, these semantic pointers will take on the following mathematical description:

$$SP = E_1 \circledast Label_{E_1} + \cdots + E_n \circledast Label_{E_n} \tag{1}$$

where E_i is a training exemplar (i.e., a semantic pointer generated through the compression of an image percept), and $Lable_{E_i}$ denotes a vector representing the category label for E_i. The symbol $\circledast$ denotes the operation of circular convolution in all equations. The implementation, use, and learnability of this operation in spiking networks are described in Stewart et al. (2011). In the case of experiments involving background knowledge and rule-based categorization, an analogous mathematical description is applied:

$$SP = R_1 \circledast K_1 + \cdots + R_n \circledast K_n \tag{2}$$

where R_i is a vector indexing a particular rule, and K_i is a representation of the contents of this rule. By sequentially retrieving and applying these rules, the model is able to infer the degree to which a given input stimulus is consistent with the category description the rules encode. Overall, (1) and (2) provide representation schemes for the semantic

pointers used in our model, but it is important to note that these schemes are chosen to accommodate specific categorization experiments and are not reflective of the representational power of our framework.

The details of the action selection system are more complicated. A number of incoming connections allow the system to monitor representational states in other neural populations in the model, and a small subset of these states are associated with actions the system performs to control information flow. The system consists of a collection of neural populations anatomically mapped to basal ganglia, and the input populations corresponding to the striatum encode the similarity (i.e., dot product) between each monitored representational state and the states that trigger particular actions. The output globus pallidus internus populations connect to the thalamus, which connects back to the rest of the model, resulting in the execution of those actions that correspond to the highest encoded similarity measure at a given time. Details about the implementation and biological plausibility of this basal ganglia model of action selection can be found in Stewart, Choo, and Eliasmith (2010). Again, though, our concern is with exploiting the functionality of this system rather than with using it to account for neural data directly.

The systems that perform perceptual and inferential evaluation are best illustrated through example. In each experimental condition, the model is first presented with a vector indicating the current task, followed by a vector corresponding to a visual stimulus to be categorized. When the task vector is passed through the visual buffer, it triggers an action that updates a working memory representation of task context, which in turn determines how the stimulus vector will be processed. In the case of a perceptual categorization task, this context representation triggers an action that compares the input stimulus to a decompressed semantic pointer in the subsystem labeled "Perceptual Evaluation" in Fig. 3. Mathematically, the subsequent output of the perceptual evaluation system can be given the following description:

$$Output = SP \circledast Stimulus^{-1} \tag{3}$$

where SP is a semantic pointer of the sort described by (1), and $Stimulus^{-1}$ is the pseudo-inverse[4] of the input vector being categorized. This output is routed to the motor buffer via an action triggered by the working memory representation of the current task context. Fig. 4 illustrates this process unfolding as the model performs a task drawn from Posner and Keele (1968).

In the case of a knowledge-based categorization task, the working memory representation of the task context triggers actions that carry out a number of inferences, each of which decompresses the same semantic pointer in a slightly different way. Decompression in this context involves extracting a representation associated with a particular rule from the semantic pointer, and then comparing this extracted representation to an input stimulus by computing a dot product. More specifically, instead of convolving a semantic pointer with the pseudo-inverse of the stimulus vector, the pseudo-inverse of a rule vector, R_i, is applied: $K_i \approx SP \circledast R_i^{-1}$. Because circular convolution is only approximately reversible, the inferential evaluation subsystem contains a simple associative memory that

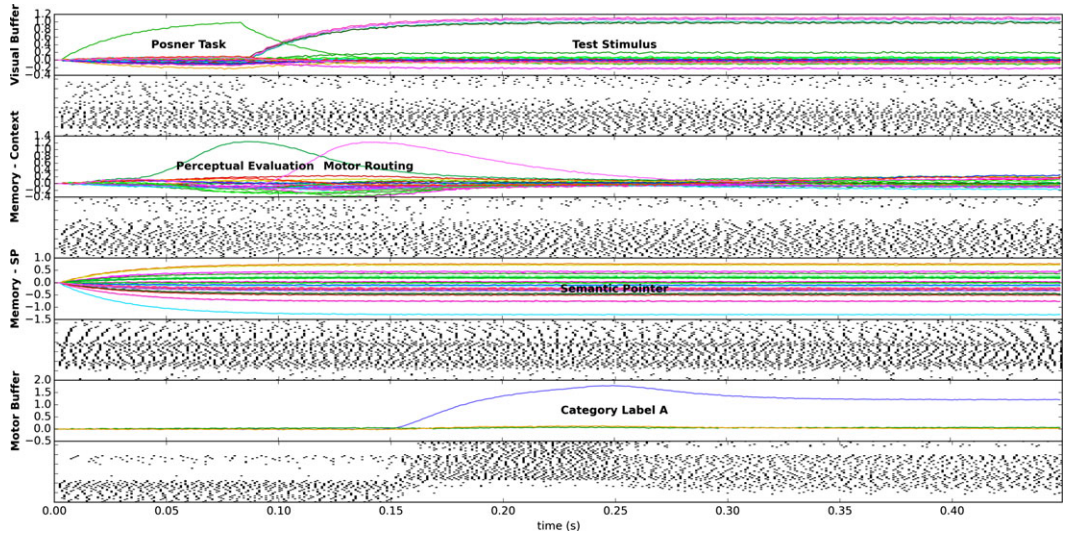

Fig. 4. An example run of the model performing a perceptual categorization task. Each plot depicts the similarity over time between the representational state in a particular neural population of the model and a set of known representational states. Below each plot is a spike raster depicting the activity in a subset of the neurons encoding each representational state. The vertical axis indicates the value of the dot product used to measure similarity, while the horizontal axis indicates time. The labels on the plot for the visual buffer, for instance, indicate that its representational state is initially similar to a vector indicating that a task from Posner and Keele (1968) is to be performed. The state then changes as the novel visual stimulus is presented, and varying degrees of similarity with known items are indicated. As the model concludes its processing, the representational state in the motor system indicates that the model has chosen "Label A" to categorize the stimulus.

"cleans up" each K_i. Then, the dot product between the input stimulus and this clean version of K_i is computed and added to a scalar value that measures the "coherence" of the stimulus with the category knowledge encoded by the semantic pointer. This running coherence measure is implemented by a neural population with recurrent connections that function to integrate input from a neural population that computes the dot product. The output of the inferential evaluation system at the conclusion of processing thus takes on the following mathematical description:

$$Output = \sum_i (SP \circledast R_i^{-1}) \cdot Stimulus \tag{4}$$

After the sequence of inferences is completed, the final state of the task context representation triggers an action that routes the output of the inferential evaluation system to the motor buffer. Fig. 5 illustrates this process unfolding as the model performs a task drawn from Lin and Murphy (1997).

We take this model to be unified in the following sense: All representations are semantic pointers, the model structure does not change across tasks, and the set of

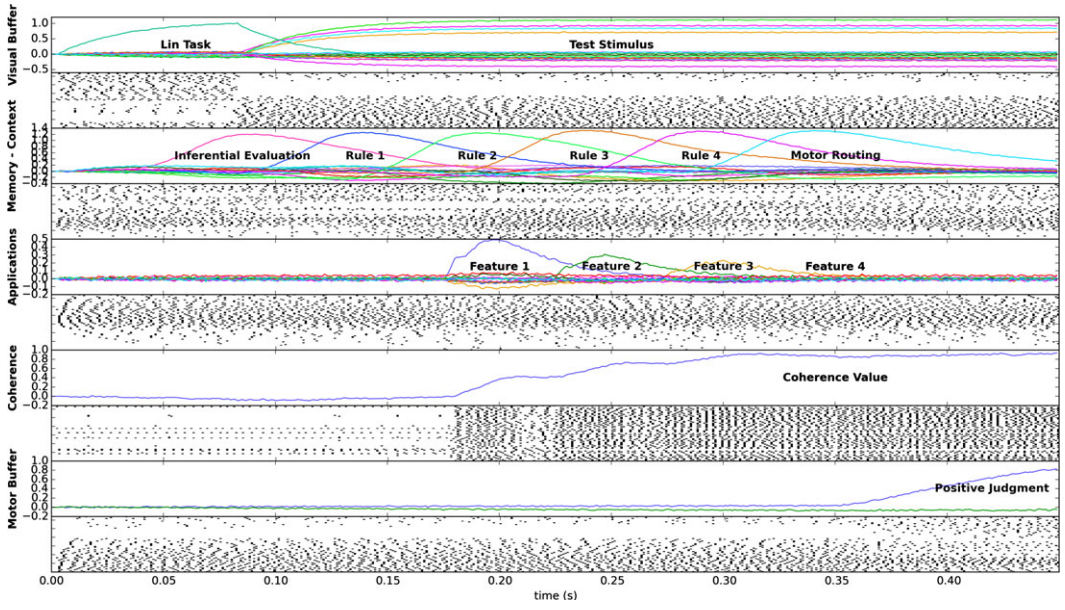

Fig. 5. An example run of the model performing an inferential categorization task. Again, each colored plot depicts the similarity over time between the representational state in a component of the model and a set of known representational states. In this case, the plot for the visual buffer indicates that its representational state is initially similar to the vector triggering a task from Lin and Murphy (1997). This task vector initiates a sequence of actions that decompress a semantic pointer stored in memory to extract a number of different rules (as labeled) for categorizing the stimulus. The effects of these rules is to assign particular weights to particular features (as shown by the values in the plot labeled "Applications"), resulting in changes to the representation that tracks the "coherence" of the stimulus with the knowledge encoded by the rules. The "Coherence" and "Applications" plots correspond to subsystems of the Inferential Evaluation system in Fig. 3.

actions that the action selection system can perform also does not change across tasks. It could be argued that because we have defined different actions for different task contexts, we have really proposed a hybrid model of some kind. However, this is analogous to arguing that a calculator does not provide a unified implementation of arithmetic. Merely changing the flow of information through the device based on input (e.g., which operation button is pushed) does not make the device implement a hybrid account of arithmetic. Representation, structure, and processing steps are held constant.

Similarly, in our model, changing the transformations performed by the action selection system is akin to manually swapping out the operation button on a calculator. We have chosen the present implementation to minimize both the complexity of the model and its run time. But, critically, changing the transformations performed by the action selection system only changes the control of information flow in the model—it does not change the nature of the representations used, the structure of the model, or the overall process by which stimuli are categorized.

7. Simulations

7.1. Prototype theory: Experiment 1

The first study we simulate is Experiment 3 of Posner and Keele's (1968) examination of dot pattern classification. The experiment was designed to investigate whether subjects abstract information about category prototypes when they are only trained to classify patterns that are generated by distorting the prototypes. In the training phase of the experiment, 30 subjects are taught through corrective feedback to categorize a set of 12 slides, each of which depicts a distinct arrangement of nine dots placed within a 30 $\times$ 30 matrix. The slides divide into three categories, and the four slides in each category are generated by randomly distorting a single "prototype" dot pattern that is not present in the training set. A distortion rule that specifies the distances each dot is moved from its starting point is used to generate the four training slides associated with each category-defining prototype. Training is considered complete when subjects are able to achieve two consecutive classifications of all 12 slides without error.

After completing of the training phase, 32 subjects are placed in a transfer phase and asked to classify a set of 24 slides without feedback. These slides consist of six old patterns (two per prototype) from the training phase, six new patterns (two per prototype) generated using the distortion rule from the training phase, six new patterns (two per prototype) generated using a weaker distortion rule, the three prototypes, and three completely random patterns. Results from the transfer phase indicate that the training patterns and prototypes are categorized best of all and equally well, while the new low-level distortion patterns and new high-level distortion patterns are categorized progressively less accurately.

To model this experiment, we first assume that all of the visual stimuli are compressed into semantic pointers using neural transformations of the sort described in Section 3. Thus, the dot patterns are presented to the model as individual vectors encoded into neural spike patterns. The three prototypes are constructed through a slightly constrained form of random vector generation to ensure a certain degree of similarity, before being normalized to unit length.[5] All vectors are 128 dimensions. To generate both the training stimuli and the transfer stimuli, the following equation is used:

$$Stimulus = Prototype + N_k(0, \sigma I) \tag{5}$$

where k refers to the dimensionality of the vector, σ refers to the level of distortion, and I refers to a $k \times k$ identity matrix. In plain language, a stimulus vector is constructed by adding a random number drawn from the normal distribution with standard deviation σ to each element of the relevant prototype vector. The value of σ is used to approximate the distortions applied to the prototype patterns by Posner and Keele. Note that the low and high distortion rules are exact ratios of one another, so a single σ value is sufficient to describe both.[6]

To run a trial of the experiment, an instance of the model is created with a semantic pointer encoding 12 labeled training images as per (1) provided as direct input into the working memory. Vectors corresponding to the task context (i.e., "Posner") and the test stimulus (e.g., "AT1" – Prototype A, Training Item 1) are then sequentially provided as input to the visual buffer. The task vector triggers an action which updates a task context representation in working memory to indicate that perceptual evaluation should be performed; this representation then triggers a further update to the task context representation, which results in the output of the perceptual evaluation system being routed to the motor buffer. Fig. 4 illustrates this process in detail for a single trial of the experiment. Each trial corresponds to 450 ms of simulated processing time.

To replicate Posner and Keele's experiment in detail, we use 32 random seeds to generate a unique instance of the model for each of the 32 experimental subjects. The seeds fix the random number generator used to set various neuron parameters in the model (e.g., maximum firing rates, preferred stimulus vector, etc.) and allow identical instances of the model to be recreated across trials. To run the complete experiment, each model instance is tested on a set of test stimuli using independent trials. The same test stimuli are used across all trials involving a particular model instance, and an overall total of $21 \times 32 = 672$ trials are conducted.[7] Results are obtained by tallying the proportion of errors the model makes in each stimulus category, and averaging this proportion over the 32 model instances.

This procedure is used in all subsequent experiments, and the same model instances are used across experiments to ensure that all results are strictly due to changes in relevant parameter values.

To evaluate the model, we found a best fit between the free parameter σ and the data reported by Posner and Keele. We perform 11 complete experiments at σ values ranging from 0.05 to 0.15, and Fig. 6 plots categorization error as function of σ for each stimulus condition. These results indicate that the model generalizes Posner and Keele's finding across a range of stimulus distortion values: Categorization accuracy is highest for the training patterns and prototypes, and it gets progressively worse for low-level and high-level distortion patterns.

Further examination of these results indicates that a σ value of 0.1 minimizes the root mean squared difference between the model results and the data. Confidence intervals of 95% for both model data and the human data are computed based on the percentages of positive and negative categorization judgments in each stimulus category.

It is helpful to conclude with a brief summary of these results. The model's performance in this experiment can be largely attributed to the mathematical structure of the semantic pointers and the way in which they are processed. To explain, each randomly generated prototype vector can be thought of as a point in 128 dimensional space, and each stimulus can be thought of as another point randomly displaced from a prototype by an amount specified by σ. As σ is increased, these displacements grow larger in both the low and high distortion conditions, and the probability of a given stimulus being located in a region of space more closely associated with an incorrect prototype (and hence an

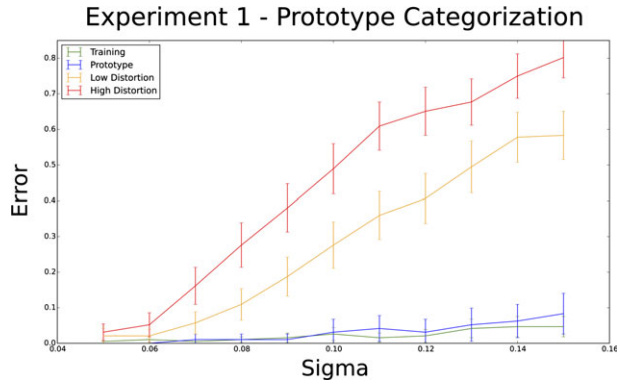

Fig. 6. Modeled error percentages for each stimulus category with varying degrees of stimulus distortion. The parameter σ is varied across 11 simulations to generate different levels of stimuli distortion, as per (5). The results indicate that larger values of σ correspond to proportionally more errors on low- and high-distortion stimuli. Errors on training and prototype stimuli, by comparison, are not significantly altered with increased values of σ. Error bars indicate 95% confidence intervals. In accordance with Posner and Keele's data, the model categorizes training items and prototypes equally well, and it makes progressively more errors on low- and high-distortion stimuli. This general pattern of results holds across a range of stimuli distortion levels, thereby generalizing Posner and Keele's observations across a range of stimuli types.

incorrect category label) increases. It is therefore not surprising that the general pattern of results displayed in Fig. 6 is obtained.

The model does, however, perform better than humans on the training patterns and prototype patterns. Two remarks can help clarify the significance of this discrepancy. First, the difference between the mean error rates in each stimulus category is roughly 10%, which corresponds to roughly one additional error in each category for every two participants in the experiment. Given that over 180 training stimuli are categorized in each experiment, this difference is actually quite small. Second, since the prototypes are randomly generated unit vectors, they can be quite dissimilar, which reduces the likelihood that miscategorization occurs, since the stimuli associated with each prototype are more likely to lie in disjoint regions of the vector space. Enforcing a minimum similarity value between the prototypes reduces the discrepancy observable in Fig. 7, but also adds a further free parameter to the model (we do not set this parameter explicitly in the results reported—see endnote 6). Adjusting the model on the basis of these factors would likely reduce the performance discrepancies observed here.

7.2. Exemplar theory: Experiment 2

While prototype theories and exemplar theories have traditionally been developed as competing explanations of the same phenomena, there are a number of experimental results that suggest that exemplar representations play a unique role in conceptual processing (Murphy, 2002). In order to account for such effects, we model an experiment (1C) by Regehr and Brooks (1993) that is designed to test the relative importance of

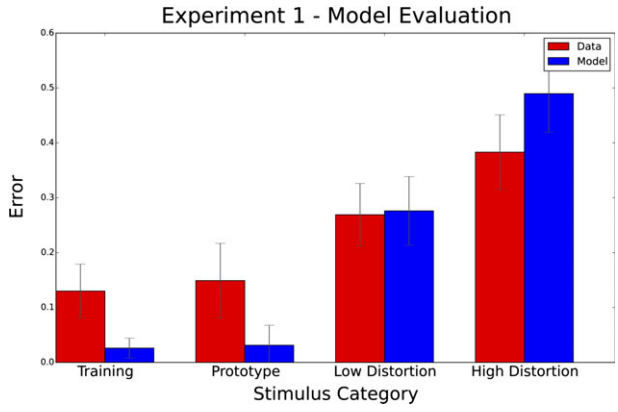

Fig. 7. Comparison of modeled and observed error percentages on a categorization task from Experiment 3, Day 1 of Posner and Keele (1968). Model results are produced through simulations employing the architecture described in Fig. 3. A σ value of 0.1 is used to generate the low distortion stimuli, and a σ level of $0.1 \times 7.7/5$ is used to generate the high distortion and training stimuli, as per the protocol of Posner and Keele. The value of σ is fit to minimize the root mean squared difference between the model results and Posner and Keele's reported results. Error bars indicate 95% confidence intervals. Like human subjects, the model categorizes the training patterns and prototypes best of all, and it makes progressively more errors on the low- and high-distortion patterns.

analytic feature matching and more holistic measures of stimulus similarity in the formation of categorization judgments. During the experiment, 32 subjects are trained to categorize drawings of imaginary creatures. Each creature possesses a unique feature set defined over five binary dimensions and belongs to one of two categories on the basis of this feature set.[8] The two categories are referred to as the "Builder" category and the "Digger" category. In order to be a Builder, a creature has to possess at least two of three specific features. Otherwise, a creature is a Digger. For each subject in the experiment, one of the following four rules is used to specify which features can be used to identify a Builder (p. 99):

1. Long legs, angular body, and spots.
2. Short legs, long neck, and spots.
3. Six legs, angular body, and spots.
4. Two legs, long neck, and spots.

The use of these separate rules is intended to balance the extent to which a given feature dimension is relevant to determining category membership (note, however, that the spots vs. no spots dimension is always relevant; Regehr & Brooks, 1993, p. 99).

Importantly, the perceptual character of each feature can vary across the drawings. For example, a "long neck" feature might have various curves in one drawing while being comparatively straight in another. The presence or absence of these sorts of perceptual differences across analytically identical drawings is used to define two experimental conditions. In the "composite" condition, analytically equivalent features are perceptually

equivalent. In the "individuated" condition, analytically equivalent features are perceptually distinct. By comparing categorization performance across these two conditions, analytic structure and perceptual similarity can be assessed for their relative importance in the formation of categorization decisions.

In the training phase of the experiment, each of the 32 subjects is taught through corrective feedback to classify a set of eight figures in accordance with one of the four rules just described. Half of the subjects are placed in the composite condition and the other half are placed in the individuated condition; each condition has its own set of eight training drawings. In the transfer phase, the subjects are asked to categorize a total of 16 drawings without feedback, eight of which are the training exemplars, and eight of which are new drawings. All of the new drawings are paired with a "twin" from the training set that differs on only one dimension (namely, the presence or absence of spots; twins are accordingly quite perceptually similar to another). New figures belonging in the same category as their twins are deemed "good transfer" (GT) items, while new figures belonging in the opposite category as their twins are deemed "bad transfer" (BT) items. The perceptual similarities between twin items can thus suggest either correct or incorrect categorization decisions: For the GT items, perceptual similarity is suggestive of the correct decision, while in the case of the BT items, perceptual similarity is suggestive of the incorrect decision.

Results from the experiment indicate that subjects in the composite condition make roughly the same percentage of categorization errors on training, GT, and BT items in the transfer phase. In the individuated condition, however, a significantly greater proportion of errors are reported for BT items.

To model the experiment, we use similar methods to those employed in the prototype simulation. We assume that the stimuli are converted into semantic pointers via a compression process, and that the structure of each semantic pointer conforms to the following mathematical description:

$$Stimulus = \sum_{F \epsilon Features} Dimension_F \circledast Value_F \tag{6}$$

where $Dimension_F$ and $Value_F$ are randomly generated vectors used to define the components of the analytic structure of each stimulus. For example, a possible dimension-value pair is *SPOTS* $\circledast$ *YES*. To increase feature individuation and approximate the difference between the composite and individuated experimental conditions, a Gaussian disturbance of variable magnitude is applied to each feature value:

$$Stimulus = \sum_{F \epsilon Features} Dimension_F \circledast (Value_F + N_k(0, \sigma I)) \tag{7}$$

where, again, k refers to the dimensionality of the vector, σ refers to the standard deviation of the Gaussian distribution, and I refers to a $k \times k$ identity matrix.

Each experimental trial is conducted using the same method employed in the Posner and Keele simulation. A semantic pointer encoding a set of eight labeled training stimuli is provided as direct input to working memory, and the visual buffer is sequentially provided with both a task vector and stimulus vector. The task vector initiates the same sequence of actions described in Fig. 4, so the change in the model's performance is only due to the use of different test stimuli and different semantic pointers. Each trial again corresponds to 450 ms of simulated processing time, and the model's categorization judgment is determined by evaluating the representational state in the model's motor system.

To assess the impact of feature individuation on categorization performance, we perform 15 experiments using stimuli generated with σ values ranging from 0.01 to 0.15. Each experiment involves testing 16 stimuli on the first 16 instances of the model, for a total of $16 \times 16 = 256$ trials. Fig. 8 plots categorization error as function of σ for each stimulus condition. Model evaluation was performed by fitting the free parameter σ to the data reported by Regehr and Brooks. We observe that a σ value of 0.02 minimizes the root mean squared difference between the model results and the data in the composite condition. Likewise, a σ value of 0.1 minimizes the root mean squared difference between the model results and the data in the individuated condition. Using different values of σ to account for the different stimulus conditions is quite reasonable, because a low σ value corresponds to comparatively small differences between analytically equivalent features on distinct stimuli, while a high σ value corresponds to comparatively large differences between such features. A direct comparison of model results and experimental data for both the composite and individuated conditions is reported in Fig. 9. Confidence intervals are computed as before.

For an intuitive explanation of these results, it is useful to again think of each stimulus as a point in high-dimensional space. The labeled training stimuli define regions in the space that are associated with one of the two category labels. When no feature individuation is present, differences in analytic structure are the *only* differences that exist between the two classes of stimuli. This is important because it means that each labeled training exemplar signals that a particular analytic structure is diagnostic of membership in a particular category. When a novel stimulus is mapped to the high-dimensional space, one can think of each training exemplar stored in memory as "voting" on the category membership of the new stimulus on the basis of points of structural overlap. For example, if a training exemplar possesses a "long neck" feature and is labeled a Builder, then the presence of this feature in the test stimulus would result in this training exemplar supplying one vote in favor of categorizing the test stimulus as a Builder. The balance of all such votes determines the resulting categorization judgment (subject to some noise given that the randomly generated vectors used in stimulus construction described in (6) are not guaranteed to be orthogonal). Overall, when no feature individuation is present, the votes produced by each memorized exemplar are sensitive only to analytic structure, which means that the change in analytic structure between GT and BT items results in a change in the number of votes a test stimulus gets for each category. This sensitivity to analytic structure explains why the model is less prone to BT error in the low feature individuation condition—the model notices the change in structure between BT items and their training pairs.

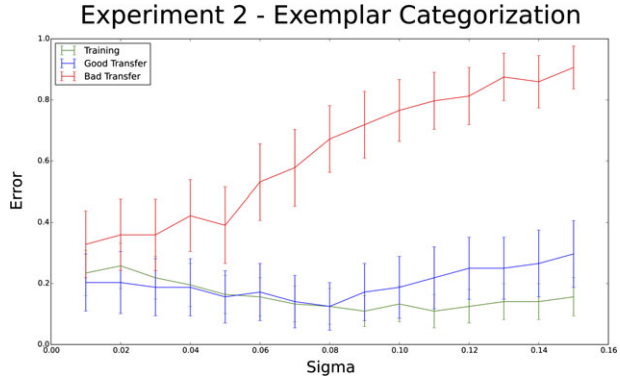

Fig. 8. Modeled error percentages in each stimulus category with varying degrees of stimulus distortion. The parameter σ is varied across 15 simulations to generate different levels of stimulus feature individuation, as per (7). The results indicate that larger values of σ correspond to proportionally more errors on bad transfer stimuli, and proportionally fewer errors on training and good transfer stimuli. Error bars indicate 95% confidence intervals. These results are consistent with the observation that highly individuated stimuli are much more likely to be misclassified when they have a twin stimulus in the opposite category (i.e., in the BT condition).

As stimuli individuation is increased, however, the votes supplied by each training stimulus become less sensitive to analytic structure. To see why, recall that feature individuation produces features that are analytically equivalent but perceptually distinct. This means that the votes supplied by a particular training exemplar only apply to test stimuli with *perceptually* similar features, since analytically equivalent features might be represented by highly distinct vectors due to the distortions used to approximate feature individuation. In this case, if a training exemplar has a highly individuated "long neck" feature, then the exemplar does *not* produce votes in favor of long necks in general. It only produces votes in favor of particular kinds of long necks. This behavior translates into increased error on BT items for the following reason. The counterpart to a BT item in the training set contains identical features on all but one dimension (Spot vs. No Spots). This means that as feature individuation is increased, the training counterpart supplies an increasingly large proportion of the votes that are relevant to categorizing the BT test stimulus. And because the training counterpart is in the *opposite* category as the BT test stimulus, these votes translate into an increased likelihood for categorization error. It is accordingly not surprising that the model performs progressively worse on BT items as the value of σ is increased over a range of experiments in Fig. 8.

Finally, it is worth remarking on the fact that the model performs better than humans on GT stimuli with minimal feature individuation. This is likely because Regehr and Brooks' additive feature rules are designed such that GT items occasionally (and uniquely) possess *none* of the features that signal membership in the opposite category. To explain, each three-feature rule can be used to divide all of the possible features into sets that are either indicative of a Builder, indicative of a Digger, or diagnostically neutral. Most stimuli, upon examination, possess one feature that is diagnostic of the category they do not

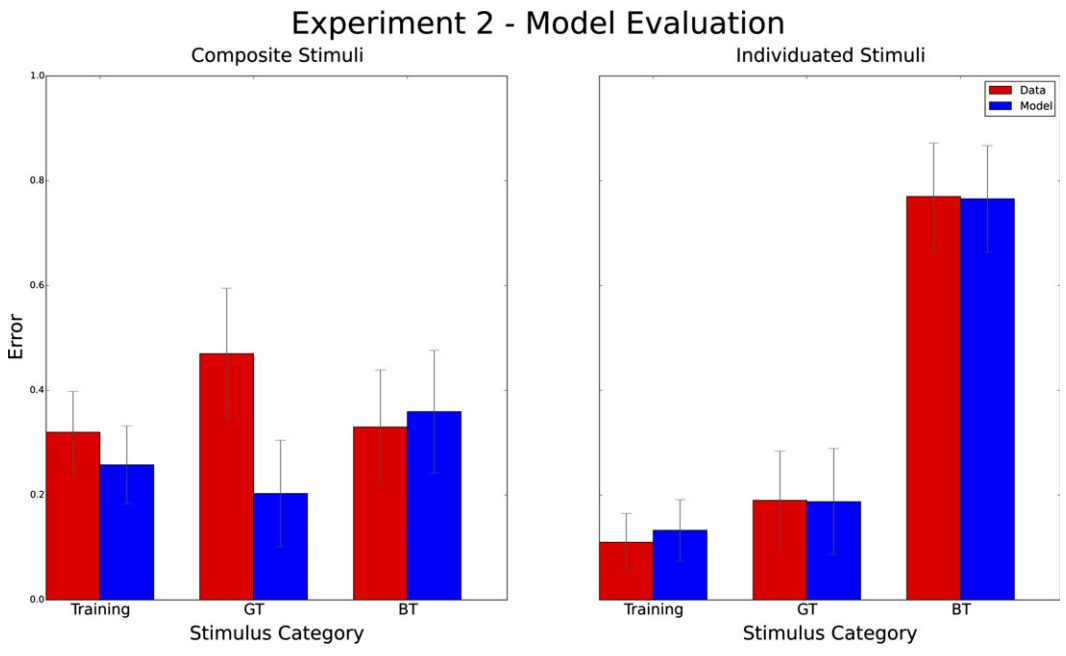

Fig. 9. Comparison of modeled and observed error percentages for composite stimuli and individuated stimuli in Experiment 1C of Regehr and Brooks (1993). The stimuli in the composite condition are generated using a σ value of 0.02, and the stimuli in the individuated condition are generated using a σ value of 0.1. These levels were selected to minimize the root mean squared difference between the model results and the data in each condition. Bad transfer (BT) items have a twin in the training set that belongs in the opposite category while differing on only one feature dimension. Good transfer (GT) items have a comparable twin item in the training set that belongs in the same category.

belong to. Half of the GT items, however, possess no features that are diagnostic of the category they do not belong to (see Regehr & Brooks, 1993, p. 102). As a result, these GT items are somewhat more likely to reside in a region of the vector space that is associated with the correct categorization judgment. It is quite possible that people are not very attuned to these subtle differences among the stimuli, especially given Regehr and Brooks' observation that many subjects report relying on only one or two features to arrive at a classification judgment. The model has no attentional mechanism that allows it place unequal priority on features, which provides a potential explanation for the more accurate performance we observe in the GT condition. Again, adjusting the model on the basis of such factors could improve the model-data fits we observe.

7.3. Theory theory: Experiment 3

In addition to the ongoing development of prototype and exemplar views, the idea that concepts are structured like intuitive theories of the categories they denote has become an increasingly popular target of research (Keil, 1989; Murphy & Medin, 1985; Rogers &

McClelland, 2004). The basic insight prompting this development is that individuals possess beliefs about things like causal relations, essences, and ontological distinctions that seem to influence how they use concepts (Keil, 1989; Murphy & Medin, 1985; Prinz, 2002; Rogers & McClelland, 2004). For example, the reason why BIRD denotes a coherent category and groups entities in the way that it does is because many of the features shared by most birds (such as flight, wings, feathers, and hollow bones) are related to one another via a set of one or more explanations: Birds can fly because they have wings, feathers, along with hollow bones; and birds fly because doing so helps them gather food and avoid predators (Murphy, 2002; Rogers & McClelland, 2004). In categorization tasks, effects of this sort manifest themselves when subjects use explanatory inferences to match an object to a category.

To provide an account of simple effects of this sort, we model Experiment 2 of Lin and Murphy (1997) in which two groups of subjects are given distinct functional descriptions of artificial categories and then asked to categorize identical sets of images. The results indicate that subjects are attentive to different features of the images depending on the category description they receive. Thus, background knowledge is shown to have an effect on subjects' performance in an image-based categorization task.

In the training phase of the experiment, 20 subjects are divided evenly into one of two groups (A and B). Each group is given a different interpretative description of a set of three training examples, each of which is comprised of four distinct features, for eight different categories. The category descriptions are devised so that of the four features present on each training example, one is characterized as functionally critical, two are characterized as functionally optional, and one is characterized as functionally irrelevant. So, in the case of the examples pictured in Fig. 10, participants in Group A are given the following description:

> Quinese hunters use tuks to catch Bondu, a type of animal that people like to eat in the Quine country. To catch a Bondu with a tuk, grab the tuk at its handle (3). Once a Bondu is spotted, throw the loop (1) over the Bondu's neck and quickly pull the string (4) at the end to tighten the loop. The cover (2) in front of the handle protects your hand from being bitten or scratched by the animal. (p. 1156)

Participants in Group B, in contrast, are given this description:

> Quinese people use tuks to spray pesticides. The triangular shaped bottle (2) contains the pesticides. When (3) is unscrewed, the pesticides flow out through the hose (4). The loop (1) is used to hang the tuk on the wall. (p. 1156)

Once a given subject learns the training examples and category descriptions for all eight categories, the subject is required to recall descriptive information about each category, and to answer questions about how best to take care of the items in each category. After the subject completes this recall process without error, they are allowed to proceed to the transfer phase of the experiment.

During the transfer phase, each subject is asked to categorize a set of new items as quickly as possible while maintaining accuracy. First, a category label (e.g., "Tuk") is presented on a computer screen for 1 s, after which an image appears. The subject then provides a Yes/No judgment before moving on to the next item. Importantly, the images that are presented after each category label vary in their consistency with the description of the category provided during the training phase. Four types of images are used. First, there are "Prototypes," which contain all four features mentioned in the category descriptions. Second, there are "Consistent A" items, which lack a feature that is functionally optional for the subjects in Group A but functionally critical for the subjects in Group B. Third, there are "Consistent B" items, which lack a feature that is functionally optional for subjects in Group B but functionally critical for subjects in Group A. Finally, there are "Control" items, which lack features that are functionally critical for subjects in both Group A and Group B.

The 20 subjects are each tested on three images per image-type for all eight of the categories. The images are presented in random order, and each subject undergoes a total of 96 trials (i.e., $3 \times 4 \times 8$). Lin and Murphy's results indicate that the prototype items elicit very high proportions of positive judgments, while the items that are consistent and inconsistent with a given subject's category knowledge elicit progressively fewer positive judgments. The control items elicit very few positive judgments.

To model this experiment, we assume that each simulated participant has learned a semantic pointer that encodes a category description as a set of simple rules. These rules function to determine whether a particular feature is important for belonging to the category under consideration. For example, one of the rules learned by a participant in Group A might be "if the item is a Tuk, then it should have a loop." The semantic pointers encoding these rules are structured in accordance with (2) as follows:

$$SP = R_1 \circledast F_1 + R_2 \circledast F_2 + R_3 \circledast F_3 + R_4 \circledast F_4 \tag{8}$$

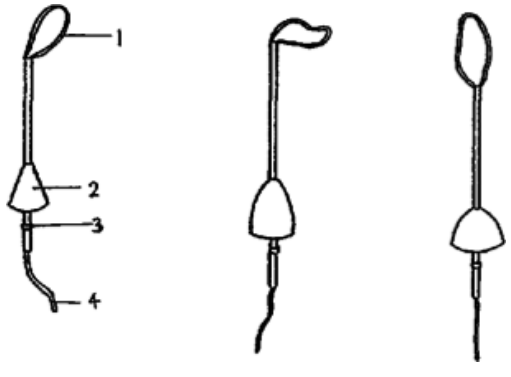

Fig. 10. Training examples reproduced from Lin and Murphy (1997). The numbers identify features of the stimuli that are given distinct functional descriptions (see text) across two experimental conditions. Once given these descriptions and the training examples, the subjects are subsequently placed in a transfer phase and asked to categorize a set of similar stimuli.

where each F_i is a semantic pointer encoding a representation of one of the features defining the category (e.g., a loop or a handle). There is accordingly one rule for each of four category-defining features. To apply a rule R_i, the action selection system performs an action that extracts F_i by decompressing the semantic pointer in the model's inferential evaluation subsystem. This decompression operation is performed by convolving the semantic pointer with the pseudo-inverse of R_i and cleaning up to a scaled version of the result. The cleanup process can be symbolized as a transition $SP \circledast R_i^{-1} \rightarrow rF_i$, where r is scalar value that defines the strength of the rule. Finally, the dot product between rF_i and the input stimulus is computed and added to a score. The value of this score after all of the rules are applied determines whether or not a positive or negative categorization judgment is routed to the motor system as output.

We do not model the process by which the values of r are learned. Rather, we assign a probability distribution to the strength of each rule, and sample from this distribution when generating each instance of the model. The distributions are calculated by assuming that on any given trial, the model should positively categorize each stimulus type with a probability equal to the proportion of positive human responses recorded by Lin and Murphy (1997) for this type. Two free parameters, a constant standard deviation σ and the mean μ_1 of distribution for the first rule, are fit to best approximate Lin and Murphy's results. The mathematics underlying this statistical technique are described in detail in Part B of the Supplemental Materials.

Fitting the parameters σ and μ_1 yields a threshold that we use to gate the output of the inferential subsystem. If the score is over the threshold, the output of the subsystem is a vector that corresponds to a positive categorization judgement. Otherwise, the output of the subsystem is a vector that corresponds to a negative categorization judgement. During each trial of the experiment, a task vector is provided as input to the visual buffer, which in turn initiates a sequence of actions that selectively decompress the semantic pointer from working memory to incrementally evaluate the stimulus. A complete illustration of this process can be found in Fig. 5.

To run a complete experiment, we create 20 instances of the model and test each instance on the same set of stimuli. For the sake of improving run-time, we only test each model instance on one stimulus per transfer type per category, for a total of $4 \times 8 \times 20 = 640$ trials. The stimulus vector in each trial is generated by adding together a set of vectors corresponding to each feature present in the stimulus type under consideration. For example, a consistent stimulus from the Tuk category would encoded as:

$$Stimulus = 1 \times F_1 + 0 \times F_2 + 1 \times F_3 + 1 \times F_4$$

where F_1, F_2, F_3, and F_4 correspond to the loop, guard, handle, and string that are referenced in the category description for Tuks. Vectors corresponding to features are randomly generated for each of the eight different categories used in the experiment. Fig. 11 reports the results of this experiment, which indicate that the model's performance is quite comparable to the human performances reported in Lin and Murphy (1997).

Overall, these results indicate that the model is able to incrementally apply knowledge encoded in a set of rules to perform rule-based stimulus categorization. When a stimulus accords with these rules, the model is very likely to judge that the stimulus belongs to the category under consideration. When a stimulus accords with the rules to a lesser degree, the model is less likely to judge that the stimulus belongs to the category under consideration. As such, the model is appropriately sensitive to how consistent a given stimulus is with respect to a simple knowledge base encoded into a set of rules. We take these results to be a very preliminary demonstration of the model's ability to account for simple rule-based knowledge effects during categorization. A discussion of possible extensions and improvements to knowledge-based conceptual processing is presented in the next section.

7.4. Model summary

We have provided a unified process model and applied it to three experimental results covering three different theories of concepts. A few remarks can help clarify what we take to be the significance of these simulations.

First, they demonstrate that semantic pointers can be manipulated and decompressed in qualitatively distinct ways. In a perceptual task, a semantic pointer is used to match the stimulus to a category label in a one-shot process that essentially amounts to pattern recognition. In an inferential task, by comparison, a semantic pointer is used to facilitate a series of computations that perform an incremental, rule-based analysis of the stimulus. Second, the simulations generalize the empirical findings reported in the studies we

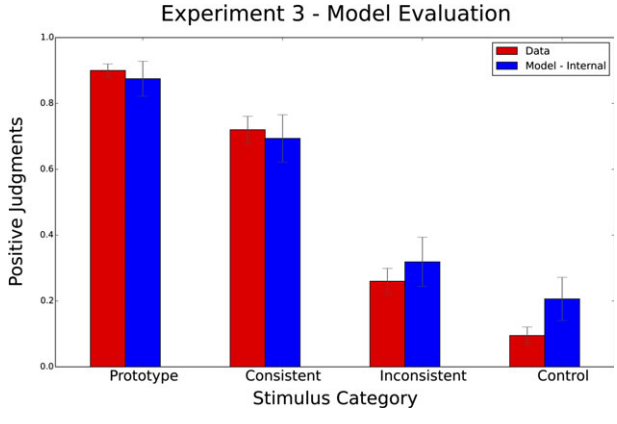

Fig. 11. Comparison of modeled and observed positive categorization judgments for an experimental task from Lin and Murphy (1997). The downward trend across the conditions indicates that as the test stimuli become more inconsistent with the relevant functional descriptions, both human subjects and the model make fewer positive category membership judgments. Error bars indicate 95% confidence intervals. The rules encoded into each semantic pointer specify the degree to which each stimulus feature is consistent with a functional description of a particular category. As such, the model's rule applications involve applying knowledge of these functional descriptions to categorize stimuli.

model. By varying a single parameter σ, we are able to replicate the patterns of perceptual categorization reported in these studies across a wide range of new stimuli. The model can therefore be used to generate novel predictions about human behavior. Third, during every experimental trial, a single semantic pointer is used to encode *all* of the category information required to arrive at a classification judgment. The model thus demonstrates how semantic pointers can offer a unified framework for studying conceptual phenomena.

One potential criticism of our model is that it fails to genuinely account for the type of phenomena that knowledge-based approaches to concepts are designed to handle. These approaches emphasize the role of causal and explanatory reasoning during categorization, and since the rules we apply in Experiment 3 essentially perform weighted feature comparisons, it seems doubtful that any such reasoning is occurring. However, rules suffice for providing explanations in the form of inferences, and they can express causal regularities when they are tied to sensory-motor manipulations (Thagard, 2012). To explain, on any appropriate understanding of reasoning, inferences can be characterized as transitions between mental states. Such transitions can be naturally captured in the form of rules or actions, and highly complex forms of reasoning involve highly complex rules and actions, while simpler forms of reasoning involve simpler rules and actions. On this understanding, it is quite apparent that our model is performing a simple form of inference. So what is really at issue is the degree rather than kind of effect being exhibited by the model. And if it is conceded that the model captures knowledge effects in *kind*, then this criticism loses much of its force.[9]

Moreover, it is simply not true that feature comparisons are inconsistent with the presence of rule-based categorization. Consider, for instance, Rips's (1989) classic study in which participants are asked to decide whether an object 3 inches in diameter is more likely to be a pizza or a quarter. Since quarters have a fixed diameter, respondents typically answer that the object is more likely to be a pizza, even though it is more similar to a quarter. This result is often taken to indicate that there are categorization processes involving rules that can override more typical processes involving assessments of similarity. Notice, though, that the relevant background knowledge about quarters can *only* be applied through a feature comparison: One must assess the diameter of the stimulus object and compare it to the known diameter of a quarter. As such, it is entirely plausible that effects of this sort could be accounted for using our model. A set of soft rules about quarters could be encoded into a semantic pointer as per Eq. (2), and one of these rules could function to very strongly penalize the coherence score of any stimulus with a sufficiently large diameter. Accounting for more sophisticated effects is, of course, an important goal for future work, but given that models of complex reasoning tasks such as the Tower of Hanoi puzzle have been implemented using semantic pointers (Stewart & Eliasmith, 2011), our claim to have provided a good starting point for handling knowledge effects is a reasonable one.

Another potential criticism of the model is that all of the explanatory insight it provides is due to the mathematical structure of the semantic pointers. While good categorization performance can be obtained using vector-based models that abstract away from neural

implementation (e.g., Knapp & Anderson, 1984), such models do not describe temporal dynamics of the sort that are present in our simulations. These dynamics can be used to make rough predictions about (a) differences in reaction times across tasks and (b) the temporal and anatomical localization of neural activity during task performance. For example, our model predicts that comparatively more neural activity would be observed in basal ganglia and thalamus during the performance of a task from Experiment 3 than during the performance of a task from Experiment 1 or 2, since more actions are performed. Similarly, the model predicts that an onset of increased motor system activity should occur later in Experiment 3 than in Experiment 1 or 2, since the action that performs motor routing is delayed as a result of the longer sequence of actions initiated by the task vector. In general, the model predicts that tasks that involve more operations on a semantic pointer should take longer than tasks that involve fewer operations, all else being equal.[10]

One final concern is that the model currently predicts no difference in reaction times (RTs) across conditions involving the same experimental task. In Experiment 3, for example, the time it takes for the model to generate a categorization judgment does not depend on the type of the stimulus being categorized (i.e., prototype and control stimuli are categorized in approximately the same amount of time). This behavior conflicts with Lin and Murphy's observation (p. 1160) of quicker response times for stimuli that are consistent with a category description in comparison to stimuli that are inconsistent with a category description. However, it is important to note that these RT differences could be due to factors we do not explicitly model. For example, it might be that the application of a rule encoded in a semantic pointer also results in processes that prime particular motor behaviors. In our model, prototype stimuli tend to achieve above-threshold coherence scores more quickly than other stimuli, even though the categorization judgment reflecting this is not routed through the motor buffer until all four rules have been decoded from the semantic pointer stored in memory. If motor priming occurs in proportion to the value of the coherence score, it might be possible to explain these RT differences. It is also worth noting that in the case of Regehr and Brook's results, no significant RT differences are observed across the stimulus conditions in experiment 1C (see p. 100). Overall, while accounting for differences across conditions in a single task is open to further exploration within our modeling framework, it is nonetheless true that the framework currently makes interesting predictions concerning the temporal and anatomical localization of neural activities across *different* tasks. We take this latter point to be the main insight offered by our modeling framework, and recognize the need for further study of reaction time data within this framework.

8. General discussion

It is worth reflecting on a few more of the general properties of the semantic pointer framework. As described, the framework offers a fairly straightforward strategy for accounting for a variety of conceptual functions. The first step in this strategy is to hypothesize the structure of the semantic pointers underlying some phenomenon of interest. The next step is to hypothesize a set of mechanisms that manipulate, compress, and

decompress these semantic pointers to bring the phenomenon about. Recent work suggests that this strategy can be used to motivate a novel cognitive architecture (Eliasmith, 2013). Recent articles have also used semantic pointers to explain priming, intentions, emotions, creativity, and consciousness (Schröder & Thagard, 2013; Schröder, Stewart, & Thagard, 2014; Thagard & Schröder, 2014; Thagard & Stewart, 2011; Thagard & Stewart, 2014).

In keeping with this breadth of application, the semantic pointer framework plausibly satisfies our five criteria for a theory of concepts. With respect to categorization, we are able to account for an important selection of experimental results and derive predictions about related results using our model. Moreover, the inferential and perceptual evaluations carried out by our model indicate that it has the capacity to explain categorization behavior involving both rules and memory in a unified manner (cf. Sloman, 1996; Smith, Patalano, & Jonides, 1998). Extensions involving additional categorization phenomena are also possible. For example, Eliasmith (2013) describes simulations in which semantic pointers are used to classify images of hand-written digits with human-level accuracy. Similarly, Hunsberger, Blouw, Bergstra and Eliasmith (2013) achieve human-level categorization performance in the tasks described in Experiments 1 and 2 while using a hierarchical visual network that takes raw images of stimuli as input.

To achieve recursive binding, we use convolution to define richly structured semantic pointers of the sort described by the representation schemes in Eqs. (1) and (2). These schemes can also be modified to account for the formation of simple natural language expressions. For example, Eliasmith (2013) demonstrates that simple sentences can be encoded using semantic pointers that bind representations of words to representations of the grammatical roles they occupy, and Stewart, Choo and Eliasmith (2014) suggest a means of parsing simple natural language sentences with this same architecture. Investigating the use of semantic pointers in natural language processing tasks is an important topic for further study.

The neural implementation criterion is satisfied through our use of LIF neurons and biologically plausible patterns of connectivity between anatomical areas such as basal ganglia, thalamus, and cortex. Additionally, since perceptual and inferential processing are performed in distinct subsystems of the model, the model is consistent with evidence indicating the abstract and concrete concepts are processed in distinct neural systems (Shallice & Cooper, 2013). This said, there are aspects of our neural implementation that require further investigation. For example, there is no direct evidence that the brain makes use of a convolution operation during conceptual processing. But given the numerous explanatory advantages accrued by postulating such an operation, we think it constitutes a reasonable working assumption.

With respect to scope, the different semantic pointers used in our model encode different kinds of representations, and it is straightforward to generalize these differences to account for highly complex and abstract concepts. For example, Eliasmith (2013) describes techniques for manipulating semantic pointers that include several hundred bound elements taken from an adult sized vocabulary, while remaining within known

anatomical constraints. In a similar vein, Crawford, Gingerich and Eliasmith (2013) have demonstrated a scalable encoding of the entire WordNet graph that employs semantic pointers.

A bit more needs to be said about representational content. Previous work suggests that the NEF is compatible with what is known as a two-factor theory of semantics (Eliasmith, 2000, 2003). Two-factor theories describe the content of a mental representation in terms of both its external causes (e.g., the stimuli that drive activity in a population of neurons) and its computational role (e.g., the subsequent effect the neural population has on other populations). So, in the case of a neural population representing, say, an auditory image, the encoding of the stimulus into neural spikes would specify the causal factor,[11] and the decoding of the spikes (to recover the signal being passed on to other neurons) would specify the computational factor. Together, these two factors define the content of the representation, and act, roughly, to pick out its extension and intension. For a semantic pointer built through the compression of numerous representations, the relevant factors would be underwritten by spiking patterns in other neural populations. The activity in these populations might, in turn, be more directly driven by perceptual stimuli, which would causally contribute to the content in the constructed semantic pointer. Of course, the theoretical details need to be fleshed out, but the general strategy of tracing functional relations among patterns of neural activity provides a principled method for identifying the content of arbitrarily complex semantic pointers.

Overall, while much work remains to be done to scale up the semantic pointer framework to account for more sophisticated conceptual phenomena, it has clear advantages over existing approaches. First, it is neurocomputationally specified to a degree that surpasses most, if not all, other accounts. Second, it offers a principled unification of a range of categorization phenomena. Third, it offers a mechanistic description of concept binding and the formation of natural language expressions. On a more philosophical front, the semantic pointer framework has tools to give an account for the wide range of different kinds of concepts and the semantic content of concepts. Discussions of semantics and scope are often ignored or bracketed in the psychological literature, while philosophers, albeit with a few exceptions (Prinz, 2002), have paid relatively little attention to empirical research on conceptual processing. Finally, we suggest that the primary contribution of this framework is that it provides a general representational scheme and biologically plausible mechanisms that can be used to implement conceptual functions often thought to be fundamentally distinct in kind (e.g., Machery, 2009). There are clear limitations to the scope of the phenomena that we have modeled here, but it should be apparent that semantic pointers provide a single representational format capable of describing all of the main *kinds* of conceptual processing. Applying this representational format to a wider range of phenomena is an important avenue for future work.

9. Conclusion

To return to our introductory remarks, we think that semantic pointers offer a promising solution to the problems framing contemporary research on concepts. Pluralism is avoided because the framework does not require the existence of multiple co-referring representational structures to account for category knowledge. In all of our simulations, one semantic pointer (in tandem with processing machinery for decompression etc.) suffices to explain all of the cognitive processes involved in using a concept to perform categorization. Since a single semantic pointer can comprehensively support conceptual processing in this manner, it makes little sense to claim that the term "concept" picks out a set of unrelated representations and processes. We suggest that if our approach is at all persuasive, concepts are here to stay.

Acknowledgments

We thank Terry Stewart for assistance with the simulations and helpful discussions on a range of topics. For other helpful suggestions, we thank the members of the University of Waterloo's Computational Neuroscience Research Group and three anonymous reviewers, whose comments have helped improve this paper in a number of ways. This research was supported by the Social Sciences and Humanties Research Council of Canada, along with the Natural Sciences and Engineering Research Council of Canada.

Notes

1. There are three other reasons for focusing on categorization simulations exclusively. First, space restrictions prohibit the modeling of numerous conceptual phenomena in one article. Second, categorization is a paradigmatic conceptual task and is therefore of interest to a wide range of researchers of different disciplinary persuasions. Third, a majority of the existing empirical research on concepts has focused on the study of categorization, and there is accordingly a rich store of data to which we can compare our results. Model-data comparisons are less feasible for tasks that have been of limited research interest.
2. The NEF defines mental representations in terms of both the encoding of stimuli into patterns of neural spikes and the decoding of sets of spike trains into the physical variables they represent (Eliasmith, 2003). To give a very simple example, two regions in the brainstem called the nuclei prepositus hypoglossi (NPH) and the vostral medial vestibular nucleus (VN) contain neurons with tuning curves that plot a relation between horizontal eye position and spiking activity (Eliasmith & Anderson, 2003, pp. 44–49). Accordingly, neurons in NPH and VN collectively "encode" a measurement of eye position into a pattern of neural spikes. The decoding

procedure involves assigning an optimal weight (either a scalar or a vector depending on the dimensionality of the decoded representation) to the responses of each neuron, and summing all such weighted responses over the relevant population and over time. In NPH and VN, the result of this sum is an estimation of the position of the eye. For a more detailed mathematical definition of these encoding and decoding relations, see Part A of the Supplemental Materials.

3. It is also worth noting that a single binding network of this sort can compute the circulation convolution of any two input vectors, and that the NEF has the resources to explain how the weights that implement such a binding network can be learned through the use of a biologically plausible Hebbian learning rule (Stewart, Bekolay, & Eliasmith, 2011).
4. Obtaining the pseudo-inverse of a vector is a simple linear transformation, and it can thus be computed on a connection between neural populations. Convolving a semantic pointer with the pseudo-inverse of a vector extracts an approximation of any item bound to that vector in the semantic pointer.
5. Specifically, the vector for the first prototype is randomly generated, while vectors for the second and third prototypes are sums of a randomly generated vector and the first vector.
6. In Posner and Keele's (1968) paper, the training and high-distortion stimuli are produced with a 7.7-bit distortion rule, and the low-distortion patterns are produced with a 5-bit distortion rule. Because the distortion values are multiples of one another, we fit the model using a "base distortion level" equal to sigma and generate the stimuli patterns in the training and high conditions using a standard deviation equal to 7.7/5 times sigma.
7. To reduce the model's runtime, no trails involving the completely random stimuli are conducted, since there is no measure of correctness for these stimuli.
8. The complete set of feature dimensions and values is as follows: body (angular or round); legs (long or short); number of legs (2 or 6); spots (yes or no); neck (long or short).
9. One further point worth making here is that it is important to distinguish between concept acquisition and concept use. In the case of the Lin and Murphy experiment, it is during the process of *concept acquisition* that knowledge of causal relations between the constriction of a loop and the killing of a Bondu is used to prioritize the presence of a loop on a Tuk. During the *use* of a concept to perform categorization, these causal relations need not be directly considered. Since we are not proposing to give a comprehensive account of concept acquisition, omitting these causal inferences from our model is reasonable.
10. There is nothing particularly significant about our choice of a 450 ms simulation window other than the fact that this window provides sufficient time for all three kinds of categorization task to complete while minimizing the overall amount of simulation time. Furthermore, there is no implicit prediction in our work that reaction times should be on the order of 450 ms, because the model excludes a

considerable amount of perceptual and motor processing that needs to be incorporated into exact estimates of reaction times.

11. This is an oversimplification. Statistical dependencies among various triggers of neural activity are used to specify the relevant causal factor to avoid problems associated with misrepresentation. See Eliasmith (2000) for details.

References

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

Barsalou, L., Simmons, W., Barbey, A., & Wilson, C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Science*, *7*(2), 84–89.

Barsalou, L., Santos, A., Simmons, W., & Wilson, C. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment and meaning* (pp. 245–283). Oxford, England: Oxford University Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

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

Crawford, E., Gingerich, M., & Eliasmith, C. (2013). Biologically plausible, human-scale knowledge representation. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 412–417). Austin, TX: Cognitive Science Society.

Dennett, D., & Viger, C. (1999). Sort-of symbols?. *Behavioral and Brain Sciences*, *22*(4), 613.

Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational conten*. PhD thesis, St. Louis: Washington University.

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy*, *C*(10), 493–520.

Eliasmith, C. (2004). Learning context sensitive logical inference in a neurobiological simulation. In S. Gayler & R. Levy (Eds.), *Compositional Connectionism in Cognitive Science*, (pp. 17–20). Menlo Park, CA: AAAI Press. AAAI Fall Symposium.

Eliasmith, C. (2013). *How to build a brain: An architecture for neurobiological cognition*. New York: Oxford University Press.

Eliasmith, C. & Anderson, C. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. New York: Oxford University Press.

Gayler, R. (1998). Multiplicative binding, representation operators and analogy. In K. D. G. Holyoak & B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences*. Sofia: NBU Press.

Georgopoulos, A., Schwartz, A., & Kettner, R. (1986). Neuronal population coding of movement direction. *Science*, *243*, 1416–1419.

Hinton, G., & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hunsberger, E., Blouw, P., Bergstra, J., & Eliasmith, C. (2013). A neural model of human image categorization. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Society* (pp. 633–638). Austin, TX: Cognitive Science Society.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. New York: Oxford University Press.

Jones, M., & Mewhort, D. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Kanerva, P. (1994). The spatter code for encoding concepts at many levels. In M. Marinaro & P. Morasso (Eds.), *Proceedings of the International Conference on Artificial Neural Networks*, (pp. 226–229). London: Springer-Verlag.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Knapp, A., & Anderson, J. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *10*(4), 616–637.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings*, (pp. 3–81). Cambridge, MA: MIT Press.

Lin, E., & Murphy, G. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology*, *23*(4), 1153–1169.

Machery, E. (2007). Concept empricism: A methodological critique. *Cognition*, *104*, 19–46.

Machery, E. (2009). *Doing without concepts*. New York: Oxford University Press.

Machery, E. (2010). Précis of doing without concepts. *Behavioral and Brain Sciences*, *33*, 195–244.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Peacocke, C. (1992). Précis of a study of concepts. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings*, (pp. 335–338). Cambridge, MA: MIT Press.

Plate, T. (2003). *Holographic reduced representations*. Stanford, CA: CSLI Publications.

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Prinz, J. (2002). *Furnishing the mind: Concepts and their perceptual bases*. Cambridge, MA: MIT Press.

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

Regehr, G., & Brooks, L. (1993). Perceptual manifestations of an analytic structure-the priority of holistic individuation. *Journal of Experimental Psychology. General*, *122*(1), 92–114.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp 21–59). New York: Cambridge University Press.

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

Rogers, T., Lambon Ralph, M., Garrard, P., Bozeat, S., McClelland, J., Hodges, J., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113–146.

Rumelhart, D., McClelland, J., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Volume *2*: Psychological and Biological Models. Cambridge, MA: MIT Press.

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review*, *120*(1), 255–280.

Schröder, T., Stewart, T., & Thagard, P. (2014). Emotions, intentions, and actions: A neural theory based on semantic pointers. *Cognitive Science*, *38*(5), 851–880.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6424–6429.

Shallice, T., & Cooper, R. (2013). Is there a semantic system for abstract words? *Frontiers in Human Neuroscience*, *7*(175), 1–10.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Smith, E., & Medin, D. (1981). *Concepts and categories*. Cambridge, MA: Harvard University Press.

Smith, E., Patalano, P., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*, 167–196.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159–217.

Solomon, K., & Barsalou, L. (2004). Perceptual simulation in property verification. *Memory and Cognition*, *32*(?), 244–259.

Stewart, T., & Eliasmith, C. (2011). Neural cognitive modeling: A biologically constrained spiking neuron model of the tower of hanoi task. In L. Carlson, C. Holscher, & T. Shipley (Ed.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 656–661). Austin, TX: Cognitive Science Society.

Stewart, T., Choo, F.-X., & Eliasmith, C. (2010). Symbolic reasoning in spiking neurons: A model of the cortex/basal ganglia/thalamus loop. In R. Catrambone, & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1100–1105). Austin, TX: Cognitive Science Society.

Stewart, T., Bekolay, T., & Eliasmith, C. (2011). Neural representations of compositional structures: Representing and manipulating vector spaces with spiking neurons. *Connection Science*, *22*(3), 145–153.

Stewart, T., Choo, F.-X., & Eliasmith, C. (2014). Sentence processing in spiking neurons: a biologically plausible left-corner parser. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 1533–1538). Austin, TX: Cognitive Science Society.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Thagard, P. (2012). *The cognitive science of science: Explanation, discovery, and conceptual change*. Cambridge, MA: MIT Press.

Thagard, P., & Schröder, T. (2014). Emotions as semantic pointers: Constructive neural mechanisms. In L. Barrett & J. Russell (Eds.), *The pyschological construction of emotions*. New York: Guilford.

Thagard, P., & Stewart, T. (2011). The aha! experience: Creativity through emergent binding in neural networks. *Cognitive Science*, *35*(1), 1–33.

Thagard, P., & Stewart, T. (2014). Two theories of consciousness: Semantic pointer competition vs. information integration. *Consciousness and Cognition*, *30*, 73–90.

Weiskopf, D. (2009). The plurality of concepts. *Synthese*, *169*, 145–173.