8 How Limited Systematicity Emerges: A Computational Cognitive Neuroscience Approach

Randall C. O'Reilly, Alex A. Petrov, Jonathan D. Cohen, Christian J. Lebiere, Seth A. Herd, and Trent Kriete

1 Introduction

In this chapter, we address the claims made by Fodor and Pylyshyn (1988) (FP88 hereafter). We strike a middle ground between classic symbolic and connectionist perspectives, arguing that cognition is less systematic than classicists claim, but that connectionist, neural-processing-based theories have yet to explain the extent to which cognition is systematic. We offer a sketch of an emerging understanding of the basis of human systematicity in terms of interactions between specialized brain systems, leveraging the computational principles identified and empirical work done in the quarter-century since the target work was published. We identify a full spectrum of processing mechanisms, arrayed along the continuum between context-sensitivity and combinatorial, systematic processing, each associated with different parts of the human brain. We find that attempting to understand the role of these different brain areas through the lens of systematicity results in a rich picture of human cognitive abilities.

FP88 make two central claims about what a classical symbol processing system must be capable of, which define a classical model:

1. *Mental representations have combinatorial syntax and semantics.* Complex representations ("molecules") can be composed of other complex representations (compositionality) or simpler "atomic" ones, and these combinations behave sensibly in terms of the constituents.
2. *Structure sensitivity of processes.* There is a separation between form and content, exemplified in the distinction between syntax and semantics, and processes can operate on the form (syntax) while ignoring the semantic content.

Taken together, these abilities enable a system to be fully *systematic* and *compositional*. Systematicity comes directly from the ability to process the form or structure of something, independent of its specific contents: if you can process sentences with a given syntax (e.g., Noun Verb Object) then you can process any constituent words in such sentences—you do not have to relearn the syntax all over again for each new word. In Chomsky's famous example, you can tell that "Colorless green ideas sleep furiously" is grammatically correct because you can encode its structural form, independent of the (lack of) meaning, while "Furiously sleep ideas green colorless" is not grammatically correct. FP88 made the point that connectionist models of that time failed to exhibit these features, and thus were insufficient models of the full power of human cognition (Fodor and Pylyshyn 1988; Fodor and McLaughlin 1990; McLaughlin 1993). This debate remains active to this day, with various critical commentaries (Aizawa 1997; Cummins 1996; Hadley 1994; Horgan and Tienson 1996; Matthews 1997; van Gelder 1990), anthologies (Macdonald and Macdonald 1995), and a book-length treatment (Aizawa 2003). Recently, Bayesian symbolic modelers have raised similar critiques of neural network models (Kemp and Tenenbaum 2008; Griffiths, Chater, Kemp, Perfors, and Tenenbaum 2010), which are defended in return (McClelland, Botvinick, Noelle, Plaut, Rogers, Seidenberg, and Smith 2010).

Qualitatively, there are two opposing poles in the space of approaches one can take in attempting to reconcile FP88 and subsequent critiques with the fact that the human brain is, in fact, made of networks of neurons. One could argue that this systematic, compositional behavior is a defining feature of human cognition, and figure out some way that networks of neurons can implement it (the "mere implementation" approach). Alternatively, one could argue that the kind of systematicity championed by FP88 is actually not an accurate characterization of human cognition, and that a closer examination of actual human behavior shows that people behave more as would be expected from networks of neurons, and not as would be expected from a classical symbol processing system (the "dismissive" approach). Few connectionist researchers have shown much enthusiasm for the project of merely implementing a symbolic system, although proof-of-concept demonstrations do exist (Touretzky 1990). Instead, there have been numerous attempts to demonstrate systematic generalization with neural networks (Bodén and Niklasson 2000; Chalmers 1990; Christiansen and Chater 1994; Hadley 1997; Hadley and Hayward 1997; Niklasson and van Gelder 1994; Smolensky 1988, 1990b; Smolensky and Legendre 2006). Also, careful examinations of language (Johnson 2004) and various aspects of human behavior have questioned whether human language, thought, and behavior really are as systematic as it is commonly assumed (van Gelder and Niklasson 1994).

An intermediate approach is to attempt to implement a symbolic system using neural networks with the intent of finding out which symbolic aspects of systematicity are plausible from a neural perspective and which are not (Lebiere and Anderson 1993). This attempt to implement the Adaptive Control of Thought—Rational (ACT-R) cognitive architecture using standard neural network constructs such as Hopfield networks and feedforward networks resulted in a considerable simplification of the architecture. This included both the outright removal of some of its most luxuriant symbolic features as neurally implausible, such as chunks of information in declarative memory that could contain lists of items and production rules that could perform arbitrarily complex pattern-matching over those chunks. More fundamentally, neural constraints on the architecture led to a modular organization that combines massive parallelism within each component (procedural control, declarative memory, visual processing, etc.) with serial synchronization of information transfers between components. That organization in turn has been validated by localization of architectural modules using neural imaging techniques (Anderson 2007). In general, this hybrid approach has resulted in an architecture that largely preserves the systematicity of the original one while greatly improving its neural plausibility. It should be pointed out, though, that systematicity in ACT-R is limited by both the skills and knowledge needed to perform any of the tasks in which it is demonstrated, and more fundamentally by the combination of the symbolic level with a subsymbolic level that controls every aspect of its operations (procedural action selection, information retrieval from memory, etc.)

The reason the systematicity debate has persisted for so long is that both positions have merit. In this chapter, we take a “middle way” approach, arguing that purely systematic symbol-processing systems do not provide a good description of much of human cognition, but that nevertheless there are some clear examples of people approximating the systematicity of symbol-processing systems, and we need to understand how the human brain can achieve this feat. Going further, we argue that a careful consideration of all the ways in which the human brain can support systematicity actually deals with important limitations of the pure symbol-processing approach, while providing a useful window into the nature of human cognition. From a neural mechanisms perspective, we emphasize the role that interactions between brain systems—including the more “advanced” brain areas, and specifically the prefrontal cortex/basal ganglia (PFC/BG) system—play in enabling the systematic aspects of human cognition. In so doing, we move beyond the limitations of traditional "connectionist" neural network models, while remaining committed to only considering neural mechanisms that have strong biological support.

Although the overall space of issues relevant to this systematicity debate is quite high-dimensional and complex, one very important principal component can be boiled down to a trade-off between *context-sensitivity and combinatoriality*. At the extreme context-sensitivity end of the spectrum, the system maintains a lookup table that simply memorizes each instance or exemplar, and the appropriate interpretation or response to it. Such a system is highly context sensitive, and thus can deal with each situation on a case-by-case basis, but is unable to generalize to novel situations. At the other end, the system is purely combinatorial and processes each separable feature in the input independently, without regard for the content in other feature channels. Such a purely combinatorial system will readily generalize to novel inputs (as new combinations of existing features), but is unable to deal with special cases, exceptions, or any kind of nonlinear interactions between features. It seems clear that either extreme is problematic and that we need a more balanced approach. This balance can be accomplished in two ways. First, one could envisage representations and information-processing mechanisms with intermediate degrees of context-sensitivity. Second, one could envisage a combination of processing systems that specialize on each of these distinct ends of the spectrum. These two strategies are not incompatible and can be combined. In this chapter, we argue that the brain incorporates functional subsystems that fall along various points of the spectrum, with evolutionarily older areas being strongly context sensitive and newer areas, notably the prefrontal cortex, being more combinatorial (though still not completely combinatorial). This limited combinatoriality is expected to produce limited systematicity in behavior. We argue that human cognition exhibits precisely this kind of limited systematicity.

The limits of human systematicity have been pointed out before (Johnson 2004; van Gelder and Niklasson 1994). Here we limit ourselves to three well-known examples from vision, language, and reasoning. Our first example is shown in figure 8.1. The context surrounding the middle letter of each word is critical for disambiguating this otherwise completely ambiguous input. A purely combinatorial system would be unable to achieve this level of context-sensitivity. Our second example is from the domain of language and illustrates the interplay between syntax and semantics. Consider the sentences:

Figure 8.1

Example of need for at least some level of context-sensitivity, to disambiguate ambiguous input in middle of each word. This happens automatically and effortlessly in people.

(1a) Time flies like an arrow.
(1b) Fruit flies like a banana.

Again, people automatically take the context into account and interpret ambiguous words such as "like" and "flies" appropriately based on this context. Our final example is from the domain of logical reasoning. Formal logic is designed to be completely context invariant and content free. Yet, psychological studies with the so-called Wason card selection task have shown that human reasoning is strongly sensitive to concrete experience. People can easily decide who to card at a bar given a rule such as "You can only drink if you are over 21," but when given the same logical task in abstract terms, their performance drops dramatically (Griggs and Cox 1982; Wason and Johnson-Laird 1972). Even trained scientists exhibit strong content effects on simple conditional inferences (Kern, Mirels, and Hinshaw 1983). More examples from other domains (e.g., the underwater memory experiments of Godden and Baddeley 1975) can easily be added to the list, but the above three suffice to illustrate the point. Human cognition is strongly context sensitive.

The standard classicist response to such empirical challenges is to refer to the competence–performance distinction (Aizawa 2003)—the idea that people are clearly capable of systematicity even if they sometimes fail to demonstrate it in particular circumstances. However, commercial symbolic AI systems are explicitly designed to have as few performance-related limitations as possible, and yet they face well-known difficulties in dealing with commonsense knowledge and practical reasoning tasks that people perform effortlessly. Arguably, these difficulties stem from the fact that a purely syntactic, formal representational system bottoms out in a sea of meaningless "atoms" and is undermined by the symbol grounding problem (Harnad 1990).

On the other hand, the classicist position also has merit. In some circumstances, it is desirable to be as context *insensitive* as possible. Perhaps the strongest examples come from the domain of deductive inference. Changing the meaning of a term halfway through a logical proof leads to the fallacy of equivocation. Consider the following fallacious argument:

(2a) A feather is light.
(2b) What is light cannot be dark.
(2c) *Therefore, a feather cannot be dark.

Here the word "light" appears in two different (context-dependent) senses in the two premises, which breaks the inferential chain. All tokens of a symbol in logic must have identical meaning throughout the proof or else the proof is not valid. Despite their natural tendency for context specificity, we can appreciate Aristotle's basic insight that the validity of deductive inference depends solely on its form and not on its content. We can learn to do logic, algebra, theoretical linguistics, and other highly abstract and formal disciplines. This fact requires explanation, just as the pervasive tendency for context-sensitivity requires explanation. Classical connectionist theories explain context-sensitivity well, but have yet to provide a fully satisfying explanation of the limited systematicity that people demonstrate.

We see the trade-off between context-sensitivity and combinatoriality as emblematic of the systematicity debate more generally. The literature is dominated by attempts to defend positions close to the extremes of the continuum. Our position, by contrast, recognizes that human cognition seems better characterized as a combination of systems operating at different points along this continuum, and for good reason: it works better that way. Thus, FP88 are extreme in advocating that human cognition should be characterized as purely combinatorial. Taken literally, the pure symbol-processing approach fails to take into account the considerable context-sensitivity that people leverage all the time that makes us truly smart, giving us that elusive common sense that such models have failed to capture all these years (and indeed Fodor himself has more recently noted that context-sensitivity of most aspects of human cognition is among the clearest and most notable findings of cognitive psychology; Fodor 2001). In other words, FP88 focus on the sharp, pristine "competence" tip of the cognitive iceberg, ignoring all the rich contextual complexity and knowledge embedded below the surface, which can be revealed in examining people's actual real-world performance. On the other side, basic 1980s-style connectionist networks are strongly weighted toward the context-sensitivity side of the spectrum, and fail to capture the considerable systematicity that people can actually exhibit, for example, when confronting novel situations or systematic domains such as syntactic processing or mathematics. For example, while McClelland and colleagues have shown that such networks can capture many aspects of the regularities and context-sensitivities of English word pronunciation (Plaut, McClelland, Seidenberg, and Patterson 1996), they also had to build into their network a precisely hand-tuned set of input features that balanced context-sensitivity and combinatoriality—in other words, the modelers, not the network, solved important aspects of this trade-off. Furthermore, such models are nowhere near capable of exhibiting the systematicity demonstrated in many other aspects of human cognition (e.g., in making grammaticality judgments on nonsense sentences, as in Chomsky's example).

As an example of the need to integrate multiple aspects of human cognition, Anderson and Lebiere (2003) proposed a test for theories of cognition called the Newell test. It consisted of a dozen criteria spanning the full range from pure combinatoriality (e.g., "behave as an almost arbitrary function of the environment") to high context-sensitivity (e.g., "behave robustly in the face of error, the unexpected, and the unknown"). They evaluated two candidate theories, ACT-R and classical connectionism, and found them both scoring well against some criteria and poorly against others. Strengths and weaknesses of the two theories were mostly complementary, indicating that human cognition falls at some intermediate point on the combinatorial–context-sensitive spectrum.

Just as we find extremism on the context-sensitivity versus combinatoriality dimension to be misguided, we similarly reject extremist arguments narrowly focused on one level of Marr's famous three-level hierarchy of computation, algorithm, and implementation. Advocates of symbol-processing models like to argue that they capture the computational level behavior of the cognitive architecture and that everything else is "mere implementation." From the other side, many neuroscientists and detailed neural modelers ignore the strong constraints that can be obtained by considering the computational and algorithmic competencies that people exhibit, which can guide top-down searches for relevant neural-processing mechanisms. We argue for a balanced view that does not single out any privileged level of analysis. Instead, we strive to integrate multiple constraints across levels to obtain a convergent understanding of human cognitive function (Jilk, Lebiere, O'Reilly, and Anderson 2008).

This convergent, multilevel approach is particularly important given our central claim that different brain areas lie at different points on the context-sensitivity versus combinatoriality continuum (and differ in other important ways as well)—the biological data (at the implementational level) provide strong constraints on the nature of the computations in these different brain areas. In contrast, a purely computational-level account of this nature would likely be underconstrained in selecting the specific properties of a larger set of specialized processing systems. Thus, most purely computational-level accounts, such as that of FP88, tend to argue strongly for a single monolithic computational-level system as capturing the essence of human cognition, whereas we argue above that such an approach necessarily fails to capture the full spectrum of human cognitive functionality.

In the following, we present a comprehensive overview of a variety of ways in which neural networks in different parts of the brain can overcome a strong bias toward context-sensitive, embedded processing that comes from the basic nature of neural processing. From both an evolutionary and online processing perspective (processing recapitulates phylogeny?), we argue that more strongly context-sensitive processing systems tend to be engaged first, and if they fail to provide a match, then progressively more combinatorial systems are engaged, with complex sequential information processing supported by the PFC/BG system providing a "controlled processing" system of last resort.

This is similar to the roles of the symbolic and subsymbolic levels in hybrid architectures such as ACT-R. The subsymbolic level is meant to replicate many of the adaptive characteristics of neural frameworks. For instance, the activation calculus governing declarative memory includes mechanisms supporting associative retrieval such as spreading activation, as well as context-sensitive pattern matching such as partial matching based on semantic similarities corresponding directly to distributed representations in neural networks. A mechanism called blending (Lebiere 1999) aggregates together individual chunks of information in a way similar to how neural networks blend together the individual training instances that they were given during learning. Together with others that similarly control procedural flow, these mechanisms constitute the highly context-sensitive, massively parallel substrate that controls every step of cognition. If they are successful in retrieving the right information and selecting the correct action, processing just flows with little awareness or difficulty (for instance, when the right answer to a problem just pops into one's head). But if they fail, then the mostly symbolic, sequential level takes over, deploying painstaking backup procedures at considerable effort to maintain the proper context information and select the right processing step at each moment.

Our most systematic, combinatorial computational model of this PFC/BG system demonstrates how an approximate, limited form of indirect variable binding can be supported through observed patterns of interconnectivity among two different PFC/BG areas (Kriete, Noelle, Cohen, and O'Reilly submitted). We have shown that this model can process items in roles they have never been seen in before, a capability that most other neural architectures entirely fail to exhibit. We then argue how this basic indirection dynamic can be extended to handle limited levels of embedding and recursion, capabilities that appear to depend strongly on the most anterior part of the PFC (APFC or frontopolar PFC, BA10; Christoff, Prabhakaran, Dorfman, Zhao, Kroger, Holyoak, and Gabrieli 2001; Bunge, Helskog, and Wendelken 2009; Koechlin, Ody, and Kouneiher 2003; Stocco, Lebiere, O'Reilly, and Anderson 2012). Thus, overall, we identify a full spectrum of processing mechanisms, arrayed along the continuum between context-sensitivity and combinatorial, systematic processing, and associated with different parts of the human brain. We find that attempting to understand the role of these different brain areas through the lens of systematicity results in a rich picture of human cognitive abilities.

2 Biological Neural Network Processing Constraints

Neuroscience has come a very long way in the intervening years since Fodor and Pylyshyn's (1988) seminal article. Yet, fundamentally, it has not moved an inch from the core processing constraints that were understood at that time and captured in the first generation of neural network models. What has changed is the level of detail and certainty with which we can assert that these constraints hold. Fundamentally, information processing in the neocortex takes place through weighted synaptic connections among neurons that adapt through local activity-dependent plasticity mechanisms. Individual pyramidal neurons in the neocortex integrate roughly 10,000 different synaptic inputs, generate discrete action potential spikes, and send these along to a similar number of downstream recipients, to whom these hard-won spikes are just a tiny drop in a large bucket of other incoming spikes. And the process continues, with information flowing bidirectionally and being regulated through local inhibitory interneurons, helping to ensure things do not light up in an epileptic fit.

Somehow, human information processing emerges from this very basic form of neural computation. Through amazing interventions like the ZIP molecule (Shema, Haramati, Ron, Hazvi, Chen, Sacktor, and Dudai 2011), which resets the learned arrangement of excitatory synaptic channels (and many other convergent experiments), we know with high confidence that learning and memory really do boil down to these simple local synaptic changes. Just as the early neural network models captured, processing and memory are truly integrated into the same neural substrate. Indeed, everything is distributed across billions of neurons and trillions of such synapses, all operating in parallel. These basic constraints are not in dispute by any serious neuroscientist working today.

The implications of this computational substrate favor context-sensitive, embedded processing, in contrast to the pure combinatoriality of the symbol processing paradigm. First, neurons do not communicate using symbols, despite the inevitable urge to think of them in this way (O'Reilly 2010). Spikes are completely anonymous, unlabeled, and nearly insignificant at an individual level. Thus, the meaning of any given spike is purely a function of its relationship to other spikes from other neurons, in the moment and over the long course of learning that has established the pattern of synaptic weights. In effect, neurons live in a big social network, learning slowly who they can trust to give them reliable patterns of activation. They are completely blind to the outside world, living inside a dark sea, relying completely on hearsay and murmurs to try to piece together some tiny fragment of "meaning" from a barrage of seemingly random spikes. That this network can do anything at all is miraculous, and the prime mover in this miracle is the learning mechanism, which slowly organizes all these neurons into an effective team of information-processing drones. Armed with many successful learning models and a clear connection between known detailed features of synaptic plasticity mechanisms and effective computational learning algorithms (O'Reilly, Munakata, Frank, Hazy, et al. 2012), we can accept that all this somehow manages to work.

The primary constraints on neural information processing are that each neuron is effectively dedicated to a finite pattern-detection role, where it sifts through the set of spikes it receives, looking for specific patterns and firing off spikes when it finds them. Because neurons do not communicate in symbols, they cannot simply pass a symbol across long distances among many other neurons, telling everyone what they have found. Instead, each step of processing has to rediscover meaning, slavishly, from the ground up, over time, through learning. Thus, information processing in the brain is fully embedded in dedicated systems. There is no such thing as "transparency"; it is the worst kind of cronyism and payola network, an immense bureaucracy. Everything is who you know—who you are connected to. We (at least those of us who love freedom and independence) would absolutely hate living inside our own brains.

This kind of network is fantastic for rapidly processing specific information, dealing with known situations and quickly channeling things down well-greased pathways—in other words, context-sensitive processing. However, as has been demonstrated by many neural network models (Plaut et al. 1996), exceptions, regularities, interactions, main effects—all manner of patterns can be recognized and processed in such a system, with sufficient learning.

From an evolutionary perspective, it is not hard to see why this is a favored form of information processing for simpler animals. We argue that the three more evolutionarily ancient brain structures—the basal ganglia, cerebellum, and hippocampus—all employ a "separator" processing dynamic, which serves to maximize context-sensitivity and minimize possible interference from other possibly unrelated learning experiences. In each of these areas, the primary neurons are very sparsely active, and thus tend to fire only in particular contexts. However, the most evolutionarily recent brain area, the neocortex, has relatively higher levels of neural activity, and serves to integrate across experiences and extract statistical regularities that can be combinatorially recombined to process novel situations. In prior work, the extreme context-sensitivity of the sparse representations in the hippocampus has been contrasted with the overlapping, more systematic combinatorial representations in the neocortex (McClelland, McNaughton, and O'Reilly 1995), yielding the conclusion that both of these systems are necessary and work together to support the full range of human cognition and memory functionality.

Next, we show how, against this overall backdrop of context-sensitive, embedded neural processing, information can be systematically transformed through cascades of pattern detectors, which can extract and emphasize some features, while collapsing across others. This constitutes the first of several steps toward recovering approximate symbol-processing systematicity out of the neural substrate.

3 The Systematicity Toolkit Afforded by Different Neural Systems

Here we enumerate the various cognitive-level capabilities that contribute to human systematicity and discuss how we think they are deployed to enable people to sometimes approximate combinatorial symbol processing. The crux of FP88's argument rests on the observation that people exhibit a level of systematicity that is compatible with the symbol processing model, and not with traditional connectionist models. Technically, systematicity is a relation among entities that are internal to the cognitive system. The *systematicity of representation* is a relation among certain representations, the *systematicity of inference* is a relation among the capacities to perform certain inferences, and so forth (Aizawa 2003; Johnson 2004). As these internal relations cannot be observed directly, the systematicity hypothesis can be tested only indirectly. Researchers have reached a broad consensus that *generalization*—the ability to apply existing knowledge to some kind of novel case—is the primary evidence for systematicity. As the structural overlap between the existing knowledge and the novel case can vary along a continuum, generalization comes in degrees. By implication, systematicity also comes in degrees (Hadley 1994; Niklasson and van Gelder 1994). Thus, it is counterproductive to view the systematicity debate as a dichotomous choice between two irreconcilable opposites. A more balanced view seems much more appropriate. In support of this view, the remainder of this chapter enumerates the sources of graded generalization that exist in neural networks and articulates how they contribute to the increasingly systematic patterns of generalization demonstrated by people.

3.1 Categorical Abstraction (Neocortex)

Networks of neurons, typically in the context of a hierarchical organization of representations, can learn to be sensitive to some distinctions in their inputs while ignoring others. The result is the formation of a categorical representation that abstracts over some irrelevant information while focusing on other relevant dimensions of variation. When processing operates on top of such categorical abstractions, it can be highly systematic, in that novel inputs with appropriate features that drive these categorical representations can be processed appropriately. Examples include commonsense categories ("dog," "cat," "chair," etc.), and also less obvious but important categories such as "up," "down," and so on. We know, for example, that the ventral visual stream, likely common to most mammals, systematically throws away spatial information and focuses contrasts on semantically relevant visual categorization (Ungerleider and Mishkin 1982; Goodale and Milner 1992). The abstract "symbolic" categories of small integer numbers have been demonstrated to exist in at least some form in monkeys and other animals, including in PFC recordings (Nieder, Freedman, and Miller 2002). In all of these cases, abstraction only works if an input has certain features that drive learned synaptic pathways that lead to the activation of a given abstract category representation. Thus, this form of generalization or systematicity implies a certain scope or basin of feature space over which it operates. But this can nevertheless be rather broad; "thing" and "one" are both rather severe abstractions that encompass a very broad scope of inputs. Categorical abstraction thus yields representations that can be used more systematically, since they are effectively stripped of context. Furthermore, it is possible to use top-down attentional processes to emphasize (or even create) certain feature activations in order to influence the categorization process and make it considerably more general—this is an important "hook" that the PFC can access, as we describe later.

One key limitation of abstraction is that, by definition, it requires throwing away specific information. This can then lead to confusion and "binding errors" when multiple entities are being processed, because it can be difficult to keep track of which abstraction goes with which concrete entity. For example, perhaps you know someone who tends to use very general terms like "thing" and "this" and "that" in conversations—it is easy to lose track of what such people are actually saying.

3.2 Relational Abstraction (Neocortex)

This is really a subtype of categorical abstraction, but one which abstracts out the relationship between two or more items. For example, "left of" or "above," or "heavier" are all relational abstractions that can be easily learned in neural networks, through the same process of enhancing some distinctions while collapsing across others (O'Reilly and Busby 2002; Hinton 1986). Interestingly, there is often an ambiguity between which way the relationship works (e.g., for "left of," which object is to the left and which is to the right?), which must be resolved in some way. One simple way is to have a dynamic focus of attention, which defines the "subject" or "agent" of the relationship. In any case, this relational ability is likely present in parietal spatial representations, and rats routinely learn "rules" such as "turn right" in mazes of various complexity. Indeed, it may be that motor actions, which often need to be sensitive to this kind of relational information and relatively insensitive to semantic "what" pathway information, provide an important driver for learning these relational abstractions (Regier and Carlson 2001). Once learned, these relational representations provide crucial generalizable ingredients for structure-sensitive processing: they are abstract representations of structure that can drive further abstract inferences about the structural implications of some situation, irrespective of the specific "contents." For example, a relational representation of physical support, such as "the glass is on the table" can lead to appropriate inferences for what might happen if the glass gets pushed off the table.

These inferences will automatically apply to any entity on a tablelike surface (even though it may seem that babies learn this fact purely through exhaustive, redundant enumeration at their high chairs).

We think these relational and inferential reasoning processes are present in a wide range of animals and can readily be inferred from their behavior. However, there are strong limits to how many steps of such reasoning can be chained together, without the benefits of an advanced PFC. Furthermore, the binding errors and tracking problems associated with abstract representations, described above, apply here as well. Thus, these relational abstractions support making abstract inferences about the implications of structural relationships, all at an abstract level, but it requires quite a bit of extra machinery to keep track of all the specific items entering into these relationships, and requires dereferencing the abstract inference back out to the concrete level again. Again, we see the PFC and its capacity for maintaining and updating temporary variable bindings as key for this latter ability.

3.3 Combinatorial Generalization (Neocortex)

Despite a bias toward context-sensitivity, it is possible for simple neural networks to learn a basic form of combinatoriality—to simply learn to process a composite input pattern in terms of separable, independent parts (Brousse 1993; O'Reilly 2001). These models develop "slot-based" processing pathways that learn to treat each separable element separately and can thus generalize directly to novel combinations of elements. However, they are strongly constrained in that each processing slot must learn independently to process each of the separable elements, because as described above, neurons cannot communicate symbolically, and each set of synapses must learn everything on its own from the ground up. Thus, such systems must have experienced each item in each "slot" at least a few times to be able to process a novel combination of items. Furthermore, these dedicated processing slots become fixed architectural features of the network and cannot be replicated ad hoc—they are only applicable to well-learned forms of combinatorial processing with finite numbers of independent slots. In short, there are strong constraints on this form of combinatorial systematicity, which we can partially overcome through the PFC-based indirection mechanism described below. Nevertheless, even within these constraints, combinatorial generalization captures a core aspect of the kind of systematicity envisioned by FP88, which manifests in many aspects of human behavior. For example, when we prepare our participants for a novel experimental task, we tell them what to do using words that describe core cognitive processing operations with which they are already familiar (e.g., push the right button when you see an A followed by an X, left otherwise); it is only the particular combination of the operations and stimuli that is novel. In many cases, a simple slot-based combinatorial network can capture this level of generalization (Huang, Hazy, Herd, and O'Reilly, in press).

3.4 Dynamic Gating (Basal Ganglia and PFC)

The basal ganglia (BG) are known to act as a dynamic gate on activations in frontal cortex, for example in the case of action selection, where the BG can "open up the gate" for a selected action among several that are being considered (Mink 1996). Anatomically, this gating takes place through a seemingly over-complex chain of inhibitory connections, leading to a modulatory or multiplicative disinhibitory relationship with the frontal cortex. In the PFC, this dynamic operates in the context of updating working memory representations, where the BG gating signal determines when and where a given piece of information is updated and maintained (Frank, Loughry, and O'Reilly 2001; O'Reilly and Frank 2006). In many ways, this is equivalent to a logic gate in a computer circuit, where a control channel gates the flow of information through another channel (O'Reilly 2006). It enables an important step of *content-independent* processing, as in structure-sensitive processing. Specifically, the BG gate can decide where to route a given element of content information, based strictly on independent control signals, and not on the nature of that content information. In the example shown in figure 8.2, "syntactic" form information

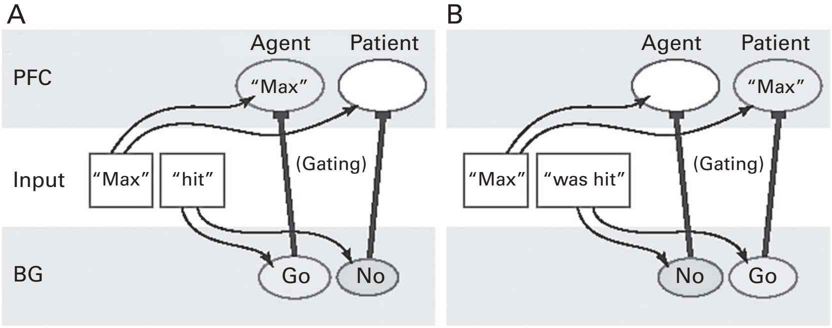

Figure 8.2
Illustration of how the basal ganglia gating dynamic with PFC can separately control the functional role assignment of other information in a content-independent fashion.

(passive vs. active verb, cued by presence or absence of keyword "was") can determine whether the preceding word is routed into an "agent" slot versus a "patient" slot in working memory. As this example makes clear, dynamic gating also helps to resolve the problem of dedicated slots for combinatorial generalization: by being able to dynamically route information into different functional slots, these slots can become more generalized, reducing the slot-explosion problem. However, it is essential to appreciate that all of this machinery must be trained up over time: the BG gating system learns through trial-and-error experience the gating strategies that lead to reward (O'Reilly and Frank 2006; Hazy, Frank, and O'Reilly 2006, 2007), and the PFC "slots" (anatomically referred to as "stripes") must learn to encode any information that they might maintain, while any other brain area that uses this maintained information must also learn to decode it (such are the basic constraints of the neural substrate, as articulated above). Thus, whatever systematicity this gating system affords must develop slowly over extensive learning experience, consistent with what we know about human symbol-processing abilities.

3.5 Active Memory Juggling and Top-Down Control (PFC/BG)

The ability to "juggle" activation states in the PFC, through the dynamic BG-mediated gating mechanism, can lead to a form of computation that escapes some of the limitations of synaptic weights (while still operating within the general confines of learning). Specifically, active maintenance plays a role like random access memory (RAM) or registers in a traditional computer architecture: whatever is being actively maintained can be rapidly updated (in a matter of a few hundreds of milliseconds), instead of requiring slow repeated learning over time. Thus, I can tell you to "pay attention to the ink color" in the ubiquitous Stroop task, and you can dynamically gate in an active representation in PFC that will drive activation of color-processing areas in posterior cortex (Herd, Banich, and O'Reilly 2006; Cohen, Dunbar, and McClelland 1990). Then, on the very next trial, you can immediately alter your behavior by gating in a "word reading" PFC representation and paying attention to the letters in the word instead of the ink color. As noted above, these PFC representations themselves have to be learned slowly over time in order to have the appropriate impact on processing elsewhere in the brain, but dynamically they can be rapidly updated and deactivated, leading to a flexibility that is absent without this PFC/BG mechanism. In principle, this kind of activation-based juggling can implement an abstract "state machine" where the active state at one point in time conditions what gets updated at the next, and relatively arbitrary sequences of such state transitions can be flexibly triggered. In the ACT-R architecture, production firing serves to update the active state of buffers, which we associate with the PFC active maintenance state (Jilk et al. 2008), demonstrating the power of this activation-based state machine for arbitrary symbolic-like processing. However, relative to ACT-R, the biology of the BG and PFC place stronger constraints on the “matching conditions” and “right-hand side” buffer update operations that result from production firing, as we discuss in greater detail below. Exactly how strong these constraints are and their implications for overall processing abilities in practice largely remains to be seen, pending development of increasingly sophisticated cognitive processing models based on this PFC/BG architecture and relevant learning mechanisms.

We have started making some progress in bridging that gap by implementing a detailed neural model of how the basal ganglia can implement the ACT-R procedural module in routing information between cortical areas associated with other ACT-R modules (Stocco, Lebiere, and Anderson 2010). Because of prior factoring of neural constraints in the evolution of the ACT-R architecture, production conditions and actions had already become naturally parallelizable, leading to a straightforward neural implementation. However, the detailed neural model reflecting the specific topology and capacity of the basal ganglia has suggested new restrictions, such as on the amount of information transfer that can occur within a single production. At the symbolic level, this is accomplished by a process of variable binding that transfers information from the condition side of the production to its action side. In terms of the neural model, that variable binding is simply realized by gating neural channels between cortical areas.

3.6 Episodic Variable Binding (Hippocampus)

The hippocampus is well known to be specialized for rapidly binding arbitrary information together in the form of a *conjunctive representation*, which can later be recalled from a partial cue (Marr 1971; McClelland et al. 1995; O’Reilly 1995; O’Reilly and Rudy 2001). This is very handy for remembering where specific objects are located (e.g., where you parked your car), the names of new people you meet, and a whole host of other random associations that need to be rapidly learned. For symbol processing, this rapid arbitrary binding and recall ability can obviously come in handy. If I tell you “John loves Mary,” you can rapidly bind the relational and abstract categorical representations that are activated, and then retrieve them later through various cues (“who loves Mary?” “John loves who?”). If I go on and tell you some other interesting information about Mary ("Mary was out last night with Richard") then you can potentially start encoding and recalling these different pieces of information and drawing some inferences, while not losing track of the original facts of the situation. However, hippocampal episodic memory also has limitations—it operates one memory at a time for both encoding and retrieval (which is a consequence of its voracious binding of all things at once), and it can take some work to avoid interference during encoding, and generate sufficiently distinct retrieval cues to get the information back out. But there is considerable evidence that people make extensive use of the hippocampus in complex symbolic reasoning tasks—undoubtedly an important learned skill that people develop is this ability to strategically control the use of episodic memory. Specific areas of PFC are implicated as these episodic control structures, including medial areas of the most anterior portion of PFC (Burgess, Dumontheil, and Gilbert, 2007).

3.7 Indirection-Based Variable Binding (PFC/BG)

The final, somewhat more speculative specialization we describe has the greatest power for advancing the kind of systematicity envisioned by FP88. By extending the basic BG dynamic gating of PFC in a set of two interconnected PFC areas, it is possible to achieve a form of *indirection* or representation by (neural) address, instead of representing content directly (Kriete et al. submitted) (figure 8.3). Specifically, one set of PFC stripes (region A) can encode a pattern of activity that drives gating in the BG for a different set of PFC stripes (region B); region A can then act as a "puppet master," pulling the strings for when the information contained in region B is accessed and updated. This then allows region A to encode the structural form of some complex representation (e.g., Noun, Verb, and Object roles of a sentence), completely independent of the actual content information that fills these structural roles (which is encoded in the stripes in region B). Critically, Kriete et al. showed that such a system can generalize in a much more systematic fashion than even networks using PFC/BG gating dynamics (which in turn generalized better than those without gating) (figure 8.4). Specifically, it was able to process a novel role filler item that had never been processed in that role before, because it had previously learned to encode the *BG address* where that content was stored. Thus, assuming that the PFC content stripes can encode a reasonable variety of information, learning only the addresses and not the contents can lead to a significant increase in the scope of generalization. Nevertheless, as in all the examples above, all of these representations must be learned slowly in

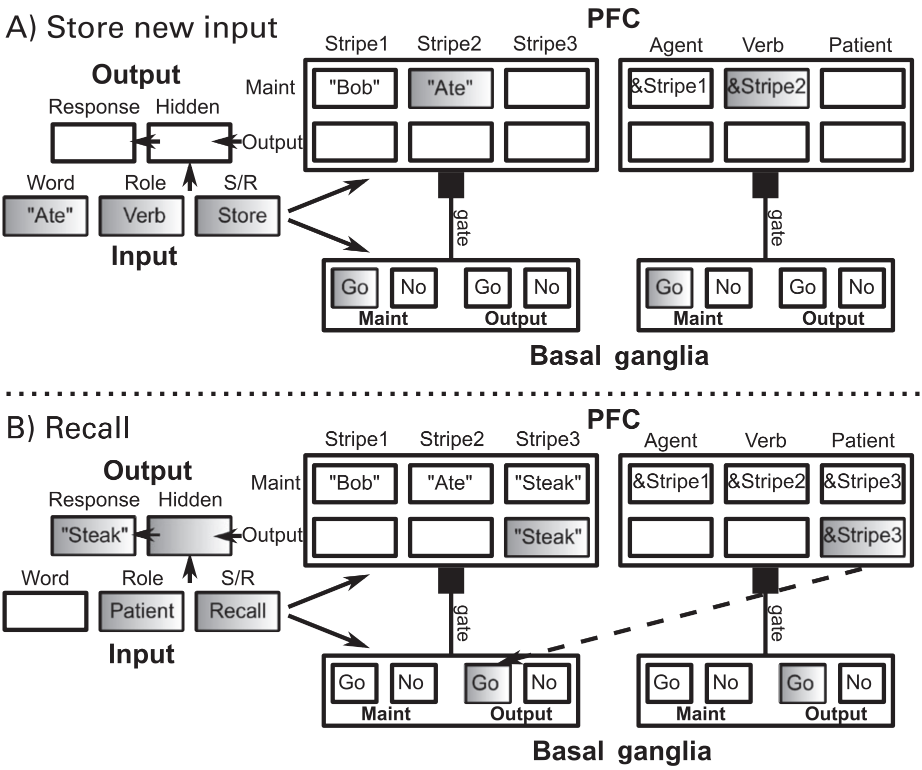

Figure 8.3

The Kriete et al. (submitted) indirection model, performing the simple sentence encoding task, demonstrating indirection in the PFC/BG working memory system. Three-word sentences are encoded one word at time, with each word associated with a role ("Agent," "Verb," or "Patient"). After encoding the sentence, the network is probed for each word using the associated roles (e.g., "What was the 'Agent' of the sentence?"). The shaded layers indicate currently active inputs. (A) One step of the encoding process for the sentence "Bob ate steak" in the PFC/BG working memory (PBWM) indirection model. The word "Ate" is presented to the network along with its current role ("Verb") and the instruction "Store" to encode this information for later retrieval. In this example, the word "Ate" is stored in Stripe2 of PFC filler stripes (left side of figure). The identity/location of Stripe2 is subsequently stored in the Verb stripe of PFC role stripes (right side of figure). The same set of events occurs for each of the other two words in the sentence (filling the agent and patient roles). (B) One step of the recall process. A role ("Patient" in the example) and the instruction "Recall" are presented as input. This drives output gating of the address information stored by that role stripe (highlighted by the dashed arrow), which in turn causes the BG units corresponding to that address to drive output gating of the corresponding filler stripe, thus outputting the contents of that stripe ("Steak").

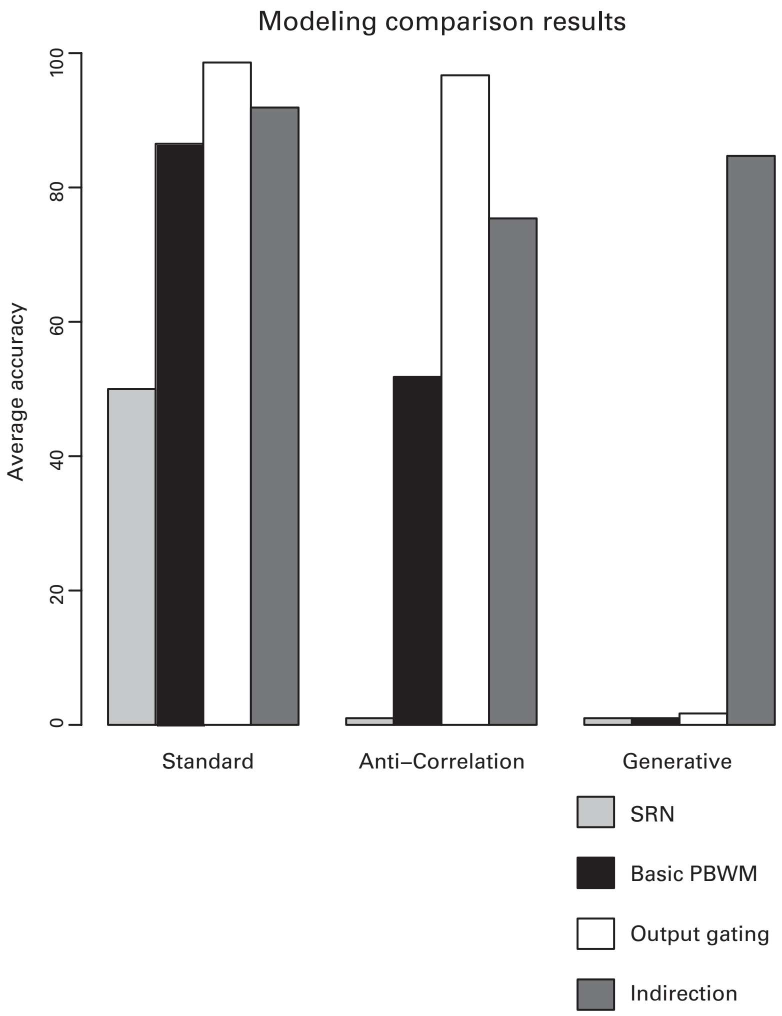

Figure 8.4
Accuracy performance of the indirection-based network juxtaposed against comparison networks, for three increasingly challenging generalization tasks. The results are grouped by task: standard, anti-correlation, and generative. Bars correspond to the four networks (from left to right): SRN, basic PBWM network with maintenance only, PBWM output gating network, and PBWMindirection network. The indirection network is the only one capable of achieving high levels of performance across all the tasks.

the first place. Our model demonstrates that, with appropriate connectivity and the same learning mechanisms used for prior PFC/BG models (O'Reilly and Frank 2006), this learning can happen naturally.

4 Putting It All Together

Having enumerated a range of different mechanisms, each of which promotes systematicity in a specific way, we now attempt to spell out some particular examples for how a complex sequence of cognitive operations, which achieves a limited approximation of classical symbol processing, could unfold through the interactions of these systems. Note that these examples are based on informed speculation, not hard data, and we do not currently have well-validated biologically based models that capture the behavior we describe. Nevertheless, we consider it plausible that this is how it is actually solved in the human brain, based on a variety of sources too numerous to explicate here. Moreover, this speculation is informed by models of similar tasks (e.g., (Lebiere 1999) in higher-level frameworks for which a correspondence to the neural architecture exists, such as ACT-R (see the section on the SAL framework below). Recently, this methodology of developing higher-level symbolic models to guide the structure and content of neural models has been applied to the complex task of sense-making (Lebiere, Pirolli, Thomson, Paik, Rutledge-Taylor, Staszewski, and Anderson submitted).

First, consider the case of multidigit mental arithmetic, for example, multiplying 42 × 17. This requires a systematic sequence of cognitive operations and keeping track of partial products, which most adults can apply to arbitrary numbers (i.e., in a fully systematic, content-independent manner). Before we consider how this happens in the general case, it is important to appreciate that if the problem was 10 × 42, for example, one would use a much faster context sensitive special-case process to arrive at the answer—people automatically and effortlessly recognize and apply these special case solutions, demonstrating the primacy of context-sensitivity as we argued above. Furthermore, in the well-studied domain of chess experts, much of the expertise is associated with this special-case pattern recognition ability and not with optimization of a fully general-purpose algorithm, whereas symbolic computer models of chess have the exact opposite profile, optimizing a general-purpose search algorithm instead of memorizing a bunch of special cases (Chase and Simon 1973).

This fundamental distinction between cognitive and algorithmic solutions arises from the hardware available to those classes of solutions.

Traditional CPUs are able to flawlessly execute billions of operations per second, but the access to the largest memory store is considerably slower and sequential. Thus, algorithmic solutions evolved to emphasize computation over memory. Neural hardware, on the other hand, is the mirror image: an excruciatingly slow and error-prone central loop (on the order of 20Hz, about eight times slower than off-the-shelf CPUs), but an extremely powerful, context-sensitive, massively parallel access to long-term memory. Cognitive solutions, therefore, evolved to emphasize memory over computation and, when computation is necessary, attempt to cache its results as efficiently and automatically as possible.

To begin on the general-case multidigit multiplication problem, people will almost certainly start by encoding the problem into hippocampal episodic memory, so they can retrieve it when interference overtakes the system and they lose track of the original problem. The next step is to recall an overall strategy for such problems, and the BG gates an abstract encoding of this strategy into an anterior portion of dorsal-lateral PFC (DLPFC). This "strategy plan" representation then activates the first step of the strategy, in a more posterior portion of DLPFC, which then drives top-down perceptual biasing in the parietal cortex to focus attention on the ones decimal place numbers (i.e., the right-most digits). Considerable categorical abstraction is required to even extract a numerical value from a particular pattern of light and dark on the retina, and abstract relational representations are required to focus on the appropriate portions of the digits, including things like top, bottom, right, and so on.

In any case, you end up activating the sub-problem of multiplying 7 × 2, which should activate the answer of 14 through well-learned parietal or perhaps temporal verbally mediated representations, perhaps even with support from the hippocampus depending on your educational status and level of recent practice. Having produced this answer, you cache away this partial product either by gating it into another available stripe in PFC (perhaps in verbal and/or numeric coding areas), or by encoding it episodically in the hippocampus (or likely both, as the hippocampus is automatically encoding everything). Next, guided by the strategic plan, you move on to the tens position in the first number, multiplying 7 × 4, encoding the 28, and so on. After each step, the partial products must be tagged and encoded in such a way that they can later be accessed for the final addition step, which in itself may require multiple substeps, with carry-overs and so on. An indirection-based variable-binding solution may be employed here, where each partial product is encoded in a different stripe, and "tagged" with the functional role of an ordinal list of items to add. Of course, items may be added incrementally in an opportunistic, context-sensitive manner, and various permutations on an overall strategy may be employed. But clearly, considerable "activation-based juggling" of information is required, along with likely several strategic hippocampal episodic encoding and retrieval steps to maintain the partial products for subsequent processing.

At some level of description, this could be considered to be a kind of classical symbol-processing system, with the hippocampus playing the role of a "tapelike" memory system in the classical Turing model and DLPFC coordinating the execution of a mental program that sequences cognitive operations over time. We do not disagree that, at that level of description, the brain is approximating a symbol-processing system. However, it is essential to appreciate that each element in this processing system has strong neurally based constraints, such that the capacity to perform this task degrades significantly with increasing number size, in a way that is completely unlike a true symbol-processing system, which can churn along on its algorithm indefinitely, putting items on the stack and popping them off at will. In contrast, the human equivalent of the "stack" is severely limited in capacity, subject to all manner of interference, and likely distributed across multiple actual brain systems. Furthermore, as noted above, the human brain will very quickly recognize shortcuts and special cases (e.g., starting with 10 × 42 as an easier problem and adjusting from there), in ways that no Turing machine would be able to. Thus, the bias toward context-sensitive processing results in very rapid and efficient processing of familiar cases—a perfectly sensible strategy for a world where environments and situations are likely to contain many of the same elements and patterns over time.

Indeed, a symbolic architecture such as ACT-R operates exactly in the way described above, with the hippocampus corresponding to declarative memory and the DLPFC corresponding to the retrieval buffer through which cognitive operations would flow for execution by the procedural system. Limitations arise through the subsymbolic level controlling the operations of the symbolic level. Chunks may exist perfectly crisp and precise at the symbolic level, but their activation ebbs and flows with the pattern of occurrence in the environment, and their retrieval is approximate, stochastic, and error-prone. Similarly, productions may implement a clocklike finite state machine, but the chaining of their individual steps into a complex processing stream is dependent on the stochastic, adaptive calculus of utilities that makes flawless execution of long procedures increasingly difficult and unlikely. Other system bottlenecks at both the architectural and subsymbolic level include limited working memory, attentional bottlenecks, and limits on execution speed for every module. Thus, hybrid symbolic-subsymbolic architectures such as ACT-R provide us with an abstraction of the capacities and limitations of neural architectures that can guide their development.

5 Discussion

We conclude with a brief discussion of some additional points of relevance to our main arguments, including the importance of data on the timecourse of learning and development on understanding the nature of human systematicity, the importance of multilevel modeling and the specific case of relating the ACT-R and Leabra modeling frameworks, and how our models compare with other related models in the literature.

5.1 The Importance of Learning and Development of Systematicity

We put a lot of emphasis on the role of "learning from the ground up" as a strong constraint on the plausibility of a given cognitive framework. Empirically, one of the strongest arguments in favor of our overall approach comes from the developmental timecourse of symbolic processing abilities in people—only after years and years of learning do we develop symbolic processing abilities, and the more advanced examples of these abilities depend critically on explicit instruction (e.g., math, abstract logic). Only in the domain of language, which nevertheless certainly is dependent on a long timecourse of exposure to and learning from a rich social world of language producers, does systematicity happen in a relatively natural, automatic fashion. And as we discuss in greater detail in a moment, language development provides many possible windows into how systematicity develops over time; it is certainly not a hallmark of language behavior right from the start.

In short, we argue that learning processes, operating over years and often with the benefit of explicit instruction, enable the development of neural dynamics involving widely distributed interacting brain systems, which support these approximate symbol-processing abilities. It is not just a matter of "resource limitations" slapped on top of a core cognitive architecture that does fully general symbol processing, as argued by FP88; rather, the very abilities themselves emerge slowly and in a very graded way, with limitations at every turn. We think this perspective on the nature of human symbolic processing argues strongly against systems that build in core symbol-processing abilities as an intrinsic part of the architecture. But unlike some of our colleagues (McClelland et al.), we nevertheless agree that these approximate symbol-processing abilities *do* develop, and that they represent an important feature that any neural network framework must account for.

One of the most famous debates between connectionists and symbol-processing advocates took place in the context of the developmental data on the U-shaped curve of overregularization of past tense morphology in English. After correctly producing irregular verbs such as "went," kids start saying things like "goed," seemingly reflecting the discovery and application of the regular "rule" ("add -ed"). First, this doesn't happen until age three or four (after considerable exposure and productive success with the language), and it is a very stochastic, variable process across kids and across time. Rates of overregularization rarely exceed a few percent. Thus, it certainly is not the kind of data that one would uphold as a clear signature of systematicity. Instead, it seems to reflect some kind of wavering balance between different forces at work in the ever-adapting brain, which we argue is a clear reflection of the different balances between context-sensitivity and combinatoriality in different brain areas. Interestingly, single-process generic neural network models do not conclusively demonstrate this U-shaped curve dynamic, without various forms of potentially questionable manipulations. Some of these manipulations were strong fodder for early critiques (Rumelhart and McClelland 1986; Pinker and Prince 1988), but even later models failed to produce this curve in a purely automatic fashion without strong external manipulations. For example, the Plunkett and Marchman (1993) model is widely regarded as a fully satisfactory account, but it depends critically on a manipulation of the training environment that is similar to the one heavily criticized by Rumelhart and McClelland (1986).

5.2 Convergent Multilevel Modeling: The SAL Framework

A valuable perspective on the nature of symbolic processing can be obtained by comparing different levels of description of the cognitive architecture. The ongoing SAL (Synthesis of ACT-R and Leabra) project provides important insight here (Jilk et al. 2008). ACT-R is a higher-level cognitive architecture that straddles the symbolic-subsymbolic divide (Anderson and Lebiere 1998; Anderson, Bothell, Byrne, Douglass, Lebiere, and Qin 2004), while Leabra is a fully neural architecture that embodies the various mechanisms described above (O'Reilly et al. 2012). Remarkably, we have found that, through different sources of constraint and inspiration, these two architectures have converged on largely the same overall picture of the cognitive architecture (figure 8.5). Specifically, both rely on the PFC/BG mechanism as the fundamental engine of cognitive sequencing from one step to the next, and this system interacts extensively with semantic and episodic declarative memory to inform and constrain the next actions selected. In ACT-R, the PFC/BG system is modeled as a production system, where production-matching criteria interrogate the contents of active memory buffers (which we associate with the PFC in Leabra). When a production fires, it results in the updating of these buffers, just as the BG updates PFC working memory in Leabra. Productions are learned through a reinforcement-based learning mechanism, which is similar across both systems.

A detailed neural model of how the topology and physiology of the basal ganglia can enable computations analog to the ACT-R production system has been developed (Stocco et al. 2010). As previously discussed, that model explains how the abstract symbolic concept of variable binding has a straightforward correspondence in terms of gating information flows between neural areas. Another major outstanding issue regarding symbolic representations is the ability to arbitrarily compose any values or structures, which in turn translates into the capacity to implement distal access to symbols (Newell 1990). The original implementation of ACT-R into neural networks (Lebiere and Anderson 1993) assumed a system of movable codes for complex chunks of information that could be decoded and their constituent parts extracted by returning to the original memory area where the composition was performed. Recent architectural developments (Anderson 2007) include the separation of the goal-related information into a goal buffer containing goal state information and an imaginal buffer containing the actual problem content. The former is associated with the working memory functionality of the prefrontal cortex whereas the latter is associated with the spatial representation and manipulation functions of the parietal cortex. This suggests that rather than using movable codes, distal access is implemented using a system of control connections that can remotely activate constructs in their original context.

5.3 Other Neural Network Approaches to Systematicity

A number of different approaches to introducing systematicity into neural network models have emerged over the years (Bodén and Niklasson 2000; Chalmers 1990; Christiansen and Chater 1994; Hadley 1997; Hadley and

R

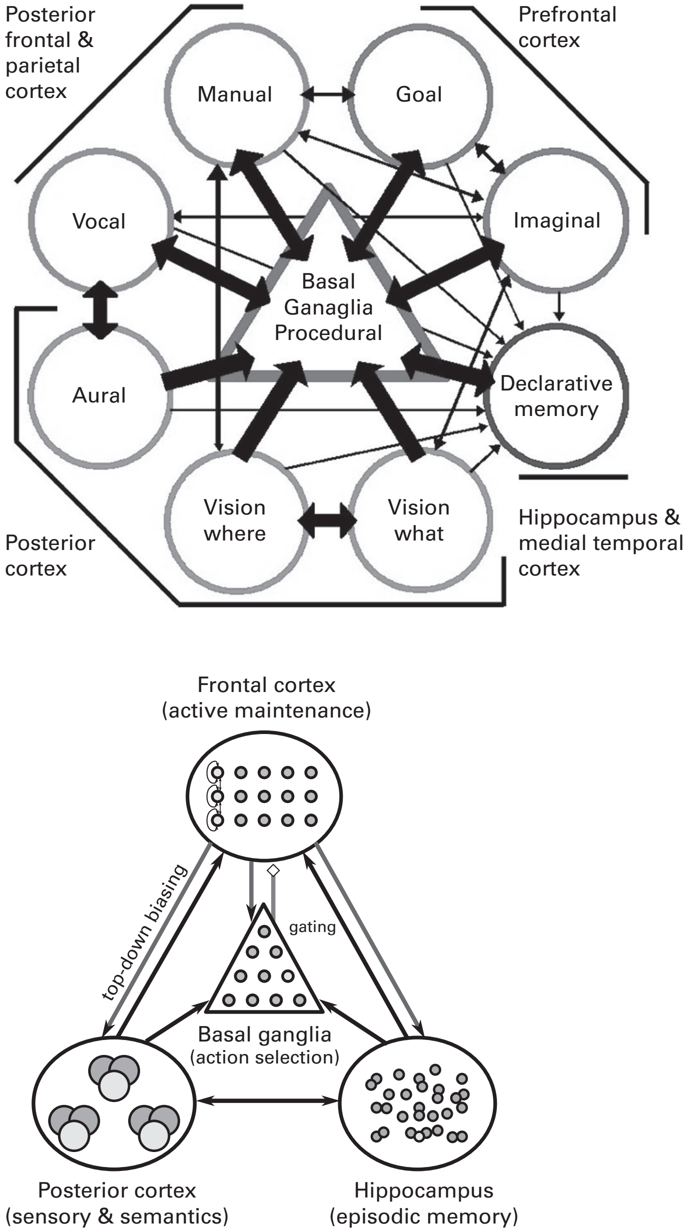

Figure 8.5

Convergent architecture between ACT-R (a) and Leabra (b) developed independently based on very different considerations.

Hayward 1997; Niklasson and van Gelder 1994; Smolensky 1988, 1990b; Smolensky and Legendre 2006). Broadly speaking, our approach is distinct from these others in focusing on a systems neuroscience perspective to the problem, both in terms of differential specializations of different brain areas and in terms of how overall symbol processing functionality can emerge through the complex interactions, over distinct time steps, between these specialized areas, as sketched above in our multidigit arithmetic example.

In terms of specific points of comparison, one of the most important mechanisms for achieving any kind of symbol processing is arbitrary variable binding, which we have argued above depends on episodic memory in the hippocampus and on the indirection-based dynamics in the PFC/BG system (Kriete et al. submitted). A number of models adopt a tensor product approach to variable binding (Plate 2008; Smolensky 1990a; Pollack 1990), which is similar in some respects to the kind of conjunctive binding achieved by the hippocampal episodic memory system. Another solution is to assume a synchrony-based binding mechanism, but we are skeptical that such a mechanism would be able to interleave multiple bindings across a phase cycle (O'Reilly and Busby 2002; O'Reilly, Busby, and Soto 2003). Furthermore, if such a mechanism were in place, it would predict a much more pervasive ability to perform arbitrary variable binding than people actually exhibit. In this respect, we think that the evidence for a long period of learning and development being required before people can even begin to demonstrate symbol-processing-like abilities is consistent with our focus on variable binding being a learned skill that involves the coordinated contributions of multiple brain areas.

As was evident in our multidigit arithmetic example, just forming a binding is only part of the problem: you also need to be able to manipulate the bound information in systematic ways. Here, we are less clear about the strong claims made by these other models: it seems that they mostly engineer various mechanisms to achieve what look to us like implementations of symbol-processing mechanisms, without a strong consideration for how such mechanisms would operate plausibly in the brain. What is conspicuously lacking is an account of how all of the complex neural processing required for these systems can be learned through experience-driven plasticity mechanisms. Our own work on this challenging problem is still in its infancy, so we certainly cannot claim to have shown how it can be learned from the ground up. Nevertheless, we remain optimistic that a learning-based approach fits best with the available human data.

6 Conclusion

After twenty-five years of earnest debate, considerable progress has been made in advancing our understanding about the nature of human systematicity. We hope that our biologically based systems neuroscience approach to these issues may provide some further insight into the nature of the human cognitive architecture and how a limited form of symbol processing can emerge through interactions between different specialized brain areas. We are excited about continuing to advance this program of research, to the point of one day showing convincingly how neural tissue can achieve such lofty cognitive functions as abstract mathematics and abstract logical reasoning.

Acknowledgments

This work was supported by NIMH grant MH079485, ONR grant N00014-13-1-0067, and the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the US Government.

References

Aizawa, K. 1997. Explaining systematicity. *Mind and Language* 12:115–136.

Aizawa, K. 2003. *The Systematicity Arguments*. New York: Kluwer Academic.

Anderson, J. R. 2007. *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

Anderson, J. R., D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. 2004. An integrated theory of the mind. *Psychological Review* 111 (4):1036–1060.

Anderson, J. R., and C. Lebiere. 1998. *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., and C. Lebiere. 2003. Optimism for the future of unified theories. *Behavioral and Brain Sciences* 26 (5):628–640.

Bodén, M., and L. F. Niklasson. 2000. Semantic systematicity and context in connectionist networks. *Connection Science* 12:111–142.

Brousse, O. 1993. *Generativity and systematicity in neural network combinatorial learning* (Technical Report CU-CS-676-93). Boulder, CO: University of Colorado at Boulder, Department of Computer Science.

Bunge, S. A., E. H. Helskog, and C. Wendelken. 2009. Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage* 46 (1):338–342.

Burgess, P. W., I. Dumontheil, and S. J. Gilbert. 2007. The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences* 11 (7):290–298.

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

Chase, W. G., and H. A. Simon. 1973. The mind's eye in chess. In *Visual Information Processing*, ed. W. G. Chase. New York: Academic.

Christiansen, M. H., and N. Chater. 1994. Generalization and connectionist language learning. *Mind and Language* 9 (3):273–287.

Christoff, K., V. Prabhakaran, J. Dorfman, Z. Zhao, J. K. Kroger, K. J. Holyoak, and J. D. Gabrieli. 2001. Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage* 14 (5):1136–1149.

Cohen, J. D., K. Dunbar, and J. L. McClelland. 1990. On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review* 97 (3):332–361.

Cummins, R. 1996. Systematicity. *Journal of Philosophy* 93 (12):591–614.

Fodor, J. 2001. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Fodor, J. A., and B. P. McLaughlin. 1990. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* 35 (2):183–204.

Fodor, J. A., and Z. W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71.

Frank, M. J., B. Loughry, and R. C. O'Reilly. 2001. Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective & Behavioral Neuroscience* 1:137–160.

Godden, D. R., and A. D. Baddeley. 1975. Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology* 66: 325–331.

R

Goodale, M. A. and D. Milner. 2010. Separate visual pathways for perception and action. *Trends in Neurosciences* 15 (1):20-25.

Griffiths, T. L., N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14 (8):357–384.

Griggs, R. A., and J. R. Cox. 1982. The elusive thematic materials effect in the Wason selection task. *British Journal of Psychology* 73:407–420.

Hadley, R. F. 1994. Systematicity in connectionist language learning. *Mind and Language* 9 (3):247–272.

Hadley, R. F. 1997. Cognition, systematicity, and nomic necessity. *Mind and Language* 12:137–153.

Hadley, R. F., and M. Hayward. 1997. Strong semantic systematicity from Hebbian connectionist learning. *Minds and Machines* 7:1–37.

Harnad, S. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42:335–346.

Hazy, T. E., M. J. Frank, and R. C. O'Reilly. 2006. Banishing the homunculus: Making working memory work. *Neuroscience* 139:105–118.

Hazy, T. E., M. J. Frank, and R. C. O'Reilly. 2007. Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 362 (1):105–118.

Herd, S. A., M. T. Banich, and R. C. O'Reilly. 2006. Neural mechanisms of cognitive control: An integrative model of Stroop task performance and fMRI data. *Journal of Cognitive Neuroscience* 18:22–32.

Hinton, G. E. 1986. Learning distributed representations of concepts. In *Proceedings of the 8th Conference of the Cognitive Science Society*, 1–12. Hillsdale, NJ: Erlbaum.

Horgan, T., and J. Tienson. 1996. *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press.

Huang, T.-R., T. E. Hazy, S. A. Herd, and R. C. O'Reilly. In press. Assembling old tricks for new tasks: A neural model of instructional learning and control. *Journal of Cognitive Neuroscience*.

Jilk, D., C. Lebiere, R. C. O'Reilly, and J. R. Anderson. 2008. SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence* 20 (3):197–218.

Johnson, K. 2004. On the systematicity of language and thought. *Journal of Philosophy* 101:111–139.

R

Kemp, C., and J. B. Tenenbaum. 2008. Structured models of semantic cognition. *Behavioral and Brain Sciences* 31:717–718.

Kern, L. H., H. L. Mirels, and V. G. Hinshaw. 1983. Scientists' understanding of propositional logic: An experimental investigation. *Social Studies of Science* 13: 136–146.

Koechlin, E., C. Ody, and F. Kouneiher. 2003. Neuroscience: The architecture of cognitive control in the human prefrontal cortex. *Science* 302:1181–1184.

Kriete, T., D. C. Noelle, J. D. Cohen, and R. C. O'Reilly. Submitted. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*.

Lebiere, C. 1999. The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft* 8:5–19.

Lebiere, C., and J. R. Anderson. 1993. A connectionist implementation of the ACT-R production system. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Lebiere, C., P. Pirolli, R. Thomson, J. Paik, M. Rutledge-Taylor, J. Staszewski, and J. Anderson. Submitted. A functional model of sensemaking in a neurocognitive architecture. *Computational Intelligence and Neuroscience*.

Macdonald, C., and G. Macdonald, eds. 1995. *Connectionism: Debates on Psychological Explanation*. Malden, MA: Blackwell.

Marr, D. 1971. Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 262 (841):23–81.

Matthews, R. 1997. Can connectionists explain systematicity? *Mind and Language* 12:154–177.

McClelland, J. L., M. M. Botvinick, D. C. Noelle, D. C. Plaut, T. T. Rogers, M. S. Seidenberg, and L. B. Smith. 2010. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14 (8): 348–356.

McClelland, J. L., B. L. McNaughton, and R. C. O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102 (3):419–457.

McLaughlin, B. P. 1993. The classicism/connectionism battle to win souls. *Philosophical Studies* 70:45–72.

Mink, J. W. 1996. The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology* 50:381–425.

R

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Nieder, A., D. J. Freedman, and E. K. Miller. 2002. Representation of the quantity of visual items in the primate prefrontal cortex. *Science* 297:1708–1711.

Niklasson, L. F., and T. van Gelder. 1994. On being systematically connectionist. *Mind and Language* 9 (3):288–302.

O'Reilly, R. 2006. Biologically based computational models of high-level cognition. *Science* 314:91–94.

O'Reilly, R. C. 1995. Biological mechanisms of controlled processing: Interactions between the prefrontal cortex and the hippocampus. Carnegie-Mellon.

O'Reilly, R. C. 2001. Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation* 13: 1199–1242.

O'Reilly, R. C. 2010. The *what* and *how* of prefrontal cortical organization. *Trends in Neurosciences* 33 (8):355–361.

O'Reilly, R. C., and R. S. Busby. 2002. Generalizable relational binding from coarse-coded distributed representations. In *Advances in Neural Information Processing Systems (NIPS) 14*, ed. T. G. Dietterich, S. Becker, and Z. Ghahramani. Cambridge, MA: MIT Press.

O'Reilly, R. C., R. S. Busby, and R. Soto. 2003. Three forms of binding and their neural substrates: Alternatives to temporal synchrony. In *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans, 168–192. Oxford: Oxford University Press.

O'Reilly, R. C., and M. J. Frank. 2006. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation* 18:283–328.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., et al. 2012. *Computational Cognitive Neuroscience*. Wiki Book, http://ccnbook.colorado.edu.

O'Reilly, R. C., and J. W. Rudy. 2001. Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review* 108:311–345.

Pinker, S., and A. Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28: 73–193.

Plate, T. A. 2008. Holographic reduced representations. *IEEE Transactions on Neural Networks* 6: 623–641.

Plaut, D. C., J. L. McClelland, M. S. Seidenberg, and K. Patterson. 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103:56–115.

Plunkett, K., and V. Marchman. 1993. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition* 48 (1):21–69.

Pollack, J. B. 1990. Recursive distributed representations. *Artificial Intelligence* 46: 77–105.

Regier, T., and L. A. Carlson. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology. General* 130:273–298.

Rumelhart, D. E., and J. L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing*, vol. 2: *Explorations in the Microstructure of Cognition: Psychological and biological models*, J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, 216–271. Cambridge, MA: MIT Press.

Shema, R., S. Haramati, S. Ron, S. Hazvi, A. Chen, T. C. Sacktor, and Y. Dudai. 2011. Enhancement of consolidated long-term memory by overexpression of protein kinase Mzeta in the neocortex. *Science* 331:1207–1210.

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–74.

Smolensky, P. 1990a. Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence* 46:159–216.

Smolensky, P. 1990b. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46 (1–2): 159–216.

Smolensky, P., and G. Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, MA: MIT Press.

Stocco, A., C. Lebiere, and J. Anderson. 2010. Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination. *Psychological Review* 117:541–574.

Stocco, A., C. Lebiere, R. C. O'Reilly, and J. R. Anderson. 2012. Distinct contributions of the caudate nucleus, rostral prefrontal cortex, and parietal cortex to the execution of instructed tasks. *Cognitive, Affective, and Behavioral Neuroscience* 12 (4):611–628.

Touretzky, D. 1990. BoltzCONS: Dynamic symbol structures in a connectionist network. *Artificial Intelligence* 46 (1–2):5–46.

Ungerleider, L. G., and Mishkin, M. 1982. Two cortical visual systems. In *The Analysis of Visual Behavior*, ed. D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, 549–586. Cambridge, MA: MIT Press.

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

van Gelder, T., and L. F. Niklasson. 1994. Classicism and cognitive architecture. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 905–909. Hillsdale, NJ: Erlbaum.

Wason, P. C., and P. N. Johnson-Laird. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.

R