SAL: an explicitly pluralistic cognitive architecture

David J. Jilk [a]; Christian Lebiere [b]; Randall C. O'Reilly [c]; John R. Anderson [b]

[a] eCortex, Inc., Boulder, Colorado, USA [b] Department of Psychology, Carnegie Mellon University, Pittsburgh,
USA [c] Department of Psychology, University of Colorado, Boulder, USA

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis
Taylor & Francis Group

TARGET ARTICLE

SAL: an explicitly pluralistic cognitive architecture

David J. Jilk[a]*, Christian Lebiere[b], Randall C. O'Reilly[c] and John R. Anderson[b]

[a]*eCortex, Inc., Boulder, Colorado, USA; [b]Department of Psychology,
Carnegie Mellon University, Pittsburgh, USA; [c]Department of Psychology,
University of Colorado, Boulder, USA*

The SAL cognitive architecture is a synthesis of two well-established constituents: ACT-R, a hybrid symbolic-subsymbolic cognitive architecture, and Leabra, a neural architecture. These component architectures have vastly different origins yet suggest a surprisingly convergent view of the brain, the mind and behaviour. Furthermore, both of these architectures are *internally* pluralistic, recognising that models at a single level of abstraction cannot capture the required richness of behaviour. In this article, we offer a brief principled defence of epistemological pluralism in cognitive science and artificial intelligence, and elaborate on the SAL architecture as an example of how pluralism can be highly effective as an approach to research in cognitive science.

Keywords: pluralism; cognitive architecture; cognitive neuroscience; cognitive models; computational models

1. Introduction

> '*The road up and down is one and the same*.'
>
> - *Heraclitus* (Freeman 1983)

One can study and describe cognition in many ways and at varying levels of detail and description. Beyond the enduring Cartesian chasm between mind and brain we have, for example, biochemists studying the details of protein channels, neuroscientists who research synaptic potentiation in individual neurons and small networks, cognitive neuroscientists who attempt to understand the behavioural function of various brain regions, and cognitive psychologists who study behaviour. Paralleling this experimental hierarchy are the simulation fields of computational neurochemistry, computational neuroscience, computational cognitive neuroscience, computational psychology, machine learning and artificial intelligence. Beyond those two scientific traditions of experimental and computational investigations lie other areas related to cognitive science, such as anthropology, education, linguistics, philosophy and human-computer interaction.

Despite the fact that all these fields in one way or another attempt to understand and/or replicate some or all of the functions of the same 1.3 kg organ, the participants rarely collaborate across boundaries. The authors believe that this separation of the larger field of cognitive science into research silos is not merely unfortunate but is likely

*Corresponding author. Email: david.jilk@e-cortex.com

to obscure the answers to the deepest theoretical questions. Newell (1973a,b) described the perils of dividing a domain into independent subdomains, and our claim might be viewed as a coarser, heterogeneous extension of those concerns. Unlike some physical sciences, cognitive science may not be amenable to this kind of reductionist arrangement of the field.

In this article, we provide a brief principled defence of pluralism in cognitive science, arguing that theories at different levels of detail and from different perspectives are mutually informative and constraining, and furthermore that no single level can capture the full richness of cognition. We then explore this perspective by providing details of two relatively successful cognitive architectures, ACT-R (Anderson and Lebiere 1998; Anderson et al. 2004; Anderson 2007) and Leabra (O'Reilly and Munakata 2000), each of which embodies theories at multiple levels of description yet has been successfully implemented as an integrated software simulation environment. Finally, we discuss the further integration of these multi-level theories into the SAL (Synthesis of ACT-R and Leabra) architecture, and show that even in early and simple tests it has demonstrated capabilities that neither of the component architectures can provide alone; and we further show that the *collaboration itself* has led to insights at the critical boundary between the architectures.

2. An argument for pluralism

Physics was perhaps the most successful field of science in the first half of the twentieth century, and consequently a great deal of recent philosophy of science centres on examples from that field. Furthermore, as a field, physics tends to have a deeply reductionist view of the world, always seeking out the smallest and most basic constituents of physical reality. It might, therefore, be surprising that pluralism is an important part of physics:

> ...psychologically we must keep all the theories in our heads, and every theoretical physicist who is any good knows six or seven different theoretical representations for exactly the same physics. He knows that they are all equivalent, and that nobody is ever going to be able to decide which one is right at that level, but he keeps them in his head, hoping that they will give him different ideas for guessing. (Feynman 1965)

Superficially this is uncontroversial. Many scientists would agree with Feynman that having different descriptions of the same phenomena helps to move the science forward and has numerous practical benefits. At a deeper level, though, most scientists (whether in physics or cognitive science) also believe that there is nevertheless *one* description that is *ontologically privileged*, the one that captures the way things *really* work. Indeed, this is what many scientists seek in their research. Consequently, a typical perspective is that, while the theories of others may be *useful*, mine is *true*. A special case of this perspective is that of the strict reductionist and his sidekick, the eliminative materialist, for whom only the *smallest* features are real, and all else is epiphenomenal.

The notion that there is an ontologically privileged description of a given phenomenon is an unsurprising outgrowth of a belief in metaphysical identity. If there is a way that things are, then there must be one right or true way to describe it. But this does not follow. A description of a phenomenon is neither the phenomenon nor an instance of it; in particular, a description is an *abstraction*. Any finite description must omit some details in favour of others that are more predictive, more revealing, more important for our current purposes. Thus for any such description, there is another that elects to incorporate some of

the omitted details and leave out others that were previously included; this second description is no less *true* – it merely has different *priorities*:

> ...one is likely to commit the common fallacy of assuming that the finer [more granular] theory is always more true than the coarser theory. The finer theory fails completely in accounting for qualitative features easily described by the coarser theory. In this case, there is an epistemic loss when one restricts oneself to the *finer* theory. Truth is here established by correspondence to different cognitive levels, each making its own contribution. It follows that *the qualitative character of the coarser theory demands recognition in its own right* despite the knowledge of the finer theory (Rohrlich 1988).

A simple concrete example of this plural validity is the behaviour of liquid water. At a fine descriptive level there are water molecules; at a much coarser level there are waves. It is easy to see the practical benefits of the two different representations. Yet reductionist instincts encourage a view that the water molecules are ontologically privileged. Alas, if one knew only of the water molecules and their micro-behaviour, one would neither predict nor expect any such phenomenon as waves. The motion of even three particles, let alone billions, is not susceptible to analytic solution, and leaving aside the computational intensity of a simulation, it is unlikely that the description and understanding of the water molecules, without a prior knowledge of waves, would be sufficiently accurate that such higher-level phenomena would arise in a simulated model. The description of water molecules instead prioritises explanations for slightly coarser phenomena such as ionic solution and cohesion. The only way that waves would arise from a simulation of water molecules is if we were to *constrain* the characterisation of the water molecules in plausible ways, find those that actually produced wave behaviour, and then determine empirically which of these constrained characterisations actually comports with the physics of the water molecules. Such an approach can provide novel predictions and refinements at both levels of description; but critically for the present discussion, it illustrates that waves and water molecules are merely two different, incomplete descriptions of a unity of physical behaviour.

In contrast to physics, psychology and artificial intelligence have sometimes had an anti-reductionist view. Although, this has abated in the past decade, with the rise of 'biologically inspired' techniques in machine learning and the use of fMRI, ERP, and single-cell recording techniques in psychology, it is still necessary to argue that the microstructure of cognition is relevant to large-scale problems such as educational issues (e.g. Anderson 2002), and machine learning practitioners are often heard to point out 'you do not need to know how birds fly to design an airplane.' At first glance this position appears to be more pragmatic than philosophical; the claim is that reductive theories are irrelevant, not unreal. But a claim that a theory is irrelevant implies that it is epiphenomenal, thus that it is ontologically inferior. The error is fundamentally the same.

It is important to see that pluralism is not the same as relativism. First, it is not the case that just *any* theory is valid: it must be consistent with the *facts*. Note that this is not the same as being consistent with all the data – no theory meets that standard, due to measurement error, experimental confounds, and the like (Kuhn 1970 pp. 146–147; Feyerabend 1993 p. 39). Second, even if there are several different descriptions of a phenomenon, all of which are consistent with the facts, it is not the case that all of them are equally *good*. Although, we intend to discredit ontological privilege, descriptions can clearly be *epistemologically privileged*. One standard for this is *parsimony*, which in very broad form can be characterised as a preference for descriptions that explain more phenomena with fewer conceptual categories; another is vertical and horizontal

coherence (Rohrlich and Hardin 1983). Furthermore, when working within a particular theoretical context, it is essential to the economy of thought to *opportunistically reify* the entities and categories of the theory. What pluralism demands is merely that one recognise this as an act of reification, rather than as an establishment of ontological privilege, and not that one must constantly give voice to alternative theoretical constructions. In essence, pluralism is an extension of instrumentalism, adding to it the claim that different systems of concepts and theories can describe the same phenomena, without contradiction, but having different aims or emphasis.

The apparently subtle distinction between ontological and epistemological privilege has major consequences for the sociology of science, and in particular for pluralism. Despite one's best efforts to see value in the work of others in the same or peripheral fields, from the perspective of ontological privilege it is difficult to avoid a thinly disguised contempt for their 'epiphenomenal' results. If instead we have a viewpoint of epistemological privilege only, then pluralism is to be embraced. Without abandoning any beliefs we may have about the underlying nature of reality, we can recognise the ineffability of those beliefs. Thus we can end what is in essence a religious war over the metaphysically true, embrace the insights and constraints generated by other approaches, and thereby hopefully enhance our own approach.

3. ACT-R

ACT-R is a cognitive architecture whose initial development was driven by modelling phenomena from the psychology laboratory. Numerous successful models have been developed for a wide range of tasks involving attention, learning, memory, problem solving, decision making and language processing. Recent years have seen a significant and relatively successful effort to embed ACT-R models into simulation environments and apply them to the performance of challenging real-world tasks. Examples of these applications include driving (Salvucci 2006), aircraft manoeuvring (Byrne and Kirlik 2005), simulated agents for computer-generated military forces (Best and Lebiere 2006), and tutoring systems of academic skills, particularly high school mathematics (Anderson and Gluck 2001).

The history of the ACT-R theory is a case study in the value of pluralism. ACT-R has its roots in the HAM (for Human Associative Memory) theory of human memory of Anderson and Bower (1973), which represented declarative knowledge as a propositional network. HAM was implemented as a running computer simulation in an attempt to handle complexity and to precisely rather than verbally specify how the model applied to the task, thus overcoming the major limitations of the mathematical theories of the 1950s and 1960s. One could view the departure from concise mathematical equations toward a collection of computational mechanisms applied to complex data structures to be a first foray into pluralism.

The next step was the introduction of the first instance of the ACT (Adaptive Control of Thought) theory, ACTE (Anderson 1976). It combined HAM's theory of declarative memory with a production system implementation of procedural memory, thus precisely specifying the process by which declarative knowledge was created and applied. Production systems were then becoming increasingly popular in cognitive science and artificial intelligence (e.g. Newell 1972, 1973a). Their combination with a theory of human memory was itself a form of pluralism by integrating mechanisms

focused on very different levels: a broad and powerful functional replication of human capabilities on the one hand, and an attention to relatively small and subtle behavioural patterns on the other. At a theoretical level, while the distinction between procedural and declarative knowledge had little support at the time, it has found increasing popularity and support from recent neuroscience evidence pointing at a dissociation between declarative and procedural memories. This ability of new neuroscience findings to illuminate long-standing debates in cognitive science is one source of our confidence in its guidance in developing our integration of the SAL architecture, as will be discussed later.

The next major step in the evolution of the ACT theory was the ACT* system (Anderson 1983), which added a neural-like calculus of activation to declarative memory that determined its functional properties. The addition of this new subsymbolic level was driven by the need to capture the soft, graded, probabilistic nature of human cognition. To each symbolic propositional node (also called a 'unit, ' and later called a 'chunk' in subsequent ACT-R versions) in memory corresponded a real-valued activation that determined its availability, from its probability of being retrieved correctly to the latency of the retrieval. The two levels are tightly integrated: an activation level is meaningless without the symbolic node to which it is attached, and a node without its activation cannot make precise quantitative behavioural predictions. This pluralistic integration of a purely symbolic cognitive theory with subsymbolic mechanisms was successful in improving correspondence to psychology laboratory results, but it also had profound architectural implications. The requirements of integration placed constraints on both the subsymbolic mechanisms (i.e. not just any set of mechanisms produce the necessary high-level behaviour) and the symbolic organisation (i.e. equivalent representations will have significantly different subsymbolic consequences), with the result that the new theory was not really a hybrid but rather a synthesis of theories of cognition at two different levels of description. Today, ACT-R modellers sometimes speak purely in terms of chunks and productions, and at other times they express the progressive changes in terms of subsymbolic quantities. The appropriate level of description depends on the modeller's immediate purposes.

Important to the history of ACT-R was the rational analysis of cognition (Anderson 1990) inspired by Marr's theory of information-processing levels (Marr 1982). The general principle of rationality states that the cognitive system operates at all times to optimise the adaptation of the behaviour of the organism. This rationality hypothesis does not imply that human cognition is perfectly optimal. Rather, it helps explain why cognition operates the way it does at the algorithmic level, given its physical limitations at the biological level and the optimum defined by the rational level that it attempts to implement. This analysis provides very strong guidance on theory development, because given a particular framework (say, an activation-based production system) it tightly constrains the set of possible mechanisms to those that satisfy the rational level. As for Marr's theory, this type of analysis is inherently pluralistic because it recognises that the same system can be analysed at different levels: from the functions that it computes to the algorithms that it uses to the details of their implementation. Moreover, it implies that a system cannot be understood at any single one of these levels but that instead the interaction between the constraints originating from each level is critical. While we initially conceived of the rational analysis as an alternative to mechanistic accounts, we later realised that the two approaches were in fact complementary. Nevertheless, many use this same framework to argue for the primacy

of the computational level and a justification for ignoring the 'implementational details' – a move that is very reasonable in the domain of computer algorithms on standard serial computers, where indeed the implementational distinctions are largely irrelevant. The recent popularity of various forms of hardware parallelism, and more exotic forms of computation such as quantum computing, have perhaps helped people appreciate that all levels really are tightly intertwined and equally important.

Constrained by this rational analysis, a new version called ACT-R (the R standing for Rational) provided a more formal basis to the subsymbolic level in terms of Bayesian statistics and extended it to procedural memory in the form of a utility calculus that determines production selection (Anderson 1993). Reflecting the growing influence of neuroscience constraints, Lebiere and Anderson (1993) attempted to implement the architecture using standard connectionist constructs. Although the resulting system, ACT-RN, was not of practical use, this connection between symbolic and connectionist levels had a fundamental impact on the development of the architecture. New mechanisms were added to capture some key connectionist properties. For instance, Lebiere, Anderson and Reder (1994) introduced a partial matching mechanism for declarative memory. This mechanism illustrated that connectionist properties can be abstracted at the subsymbolic level, for example, by reducing distributed representations to similarities that are then combined with activation to yield semantically-driven retrievals. Also, many complex symbolic constructs that were found to be too difficult to implement and thus neurally implausible were removed from the architecture. These changes resulted in a finer-grained, more constrained version of the architecture (Anderson and Lebiere 1998), which can be viewed as embodying another form of pluralism between its computational nature, its mathematical and statistical roots in the rational analysis, and the neural constraints that limited its complexity.

The most recent version of the ACT-R theory (Anderson et al. 2004) has integrated more granular theories along a different dimension, that of architectural organisation. Under the pressure of accommodating the wide range of tasks mentioned above, the architecture has added fairly detailed modules that represent perceptual attention and motor programming. To accommodate new knowledge from fMRI and other neural techniques regarding the functioning and organisation of the brain (Anderson 2007), it has adopted a highly modular structure to incorporate these new capabilities, as well as to modularise long-standing capabilities such as long-term declarative memory, goal processing and procedural competence (Figure 1). The information processing in each of these modules is largely isolated from the information processing in others. They communicate with one another by putting information into limited-capacity buffers, and production rules coordinate their action by recognising patterns in the buffers and making further requests of the modules. A major benefit of this modular approach is that the architecture has become *dynamically* pluralistic, by facilitating the integration of a variety of modules, often borrowed from other architectures. For example, development of the perceptual and motor modules (including more specifically the visual and manual modules illustrated in Figure 1) was heavily influenced by the EPIC architecture (Meyer and Kieras 1997). In addition, the modules have mappings to brain regions, and this has enabled the use of cognitive neuroscience data, particularly brain imaging, to guide the further development of models and the architecture. Once again, extending the architecture to incorporate more granular theories has provided strong constraints for productive development at both the higher and lower levels.
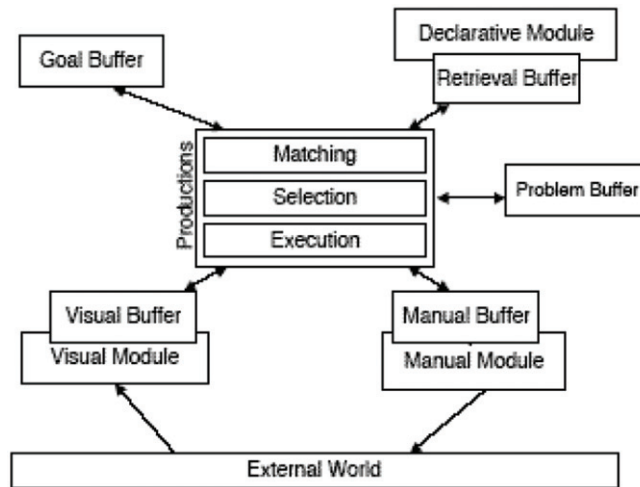
Figure 1. Overview of the ACT-R architectural organisation.

4. Leabra

Development of the Leabra framework also grew out of a desire to reproduce laboratory results in simulation, but from an entirely different starting point. Beginning with fundamental neural mechanisms, Leabra integrates into one coherent framework a set of basic neural learning and processing mechanisms (O'Reilly 1996; O'Reilly 1998; O'Reilly and Munakata 2000; O'Reilly 2001) that have been otherwise separately investigated in the neural modelling community.

In Leabra, individual neuron behaviour is governed by a point-neuron activation function that uses simulated ion channels to update a membrane potential, with a non-linear, thresholded output to other neurons. While this function is substantially simplified over detailed neurochemistry simulations, it is behaviourally richer than more abstract neural networks, and produces results that cannot be replicated in its absence (e.g. shunting inhibition, which itself is critical for balancing inhibitory competition with distributed representations, as described below). Similarly, Leabra incorporates three different learning mechanisms: small increments of Hebbian learning, a substantial component of error-driven learning, and in some cases reinforcement learning, which together produce better overall learning than any one mechanism alone (O'Reilly and Munakata 2000; O'Reilly 2001; O'Reilly and Frank 2006; O'Reilly, Frank, Hazy, and Watz 2007).

Importantly, all three mechanisms are organised to correspond to biologically plausible formulations. Hebbian and other forms of associative learning are well established as having a biological foundation (Lisman, Lichtman, and Sanes 2003). The error-driven component is based on post-synaptic calcium dynamics (Jilk, Cer and O'Reilly 2003) and bidirectional excitatory connectivity, yet provides results that are provably similar to those of backpropagation (O'Reilly 1996). The Leabra theory of reinforcement learning depends on a complex architectural arrangement called 'PVLV' that maps closely to subcortical neural structures (O'Reilly, Frank, Hazy, and Watz 2007).

Thus at its most granular level, Leabra posits a theory of neuron function and learning that is both consistent with the known biology and comparable in effectiveness with

machine learning techniques. This theory attempts to isolate those biological factors that are essential to the computational result at both a neural and network level. These mutually constraining factors from two very different fields were essential to the development of the framework.

Leabra also provides a theory of how such neurons connect and interact in networks. First, it includes both small-scale and regional inhibitory fields, which both maintain overall activity at desirable levels and induce representational specialisation of neurons and groups of neurons. Critically, the inhibition is such that multiple neurons within a functional area can be active, providing all the benefits of distributed representations, in contrast with the more prevalent single winner-takes-all (WTA) algorithms, which produce only localist representations (O'Reilly 1998; O'Reilly and Munakata 2000). Leabra's approach to inhibition is guided by the anatomy and behaviour of inhibitory interneurons in the brain, but the actual implementation in a k-winners-take-all (kWTA) function is via a computational abstraction, reflecting a pragmatic, pluralistic approach. Second, Leabra includes bidirectional connectivity among regions, which is a striking feature of the brain's anatomy. In a simulated network, bidirectional connections result in dynamic constraint satisfaction between top-down and bottom-up influences, which is critically important to performing cognitive tasks such as interpretation of ambiguous stimuli or focusing attention in relation to current goals. They also cause strong attractor states, a crucial representational feature, to develop during learning.

The large-scale architectural organisation of Leabra (Figure 2) includes three major brain systems: the posterior cortex, specialised for perceptual and semantic processing using slow, integrative learning; the hippocampus, specialised for rapid encoding of novel information using fast, arbitrary learning; and the frontal cortex/basal ganglia complex, specialised for active and flexible maintenance of goals and other context information, which serves to control or bias processing throughout the system. This latter system also incorporates various neuromodulatory systems, such as dopamine, norepinephrine, and
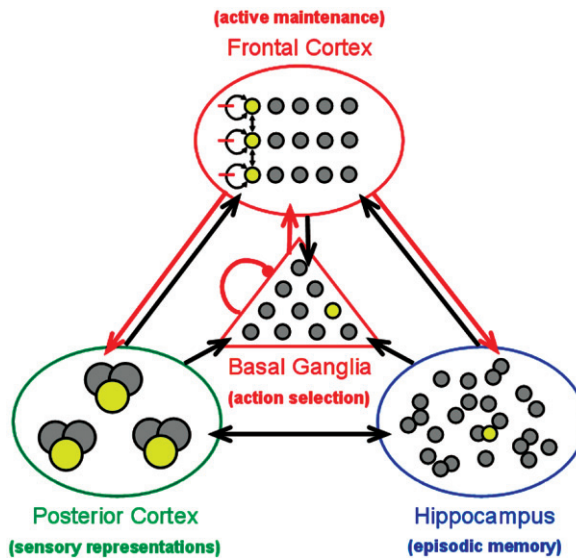

Figure 2. Overview of the Leabra architectural organisation.

acetylcholine, that are driven by cortical and subcortical areas (e.g. the amygdala, ventral tegmental area, substantia nigra pars compacta and locus ceruleus) involved in emotional and motivational processing. These neuromodulators are important for regulating overall learning and decision-making characteristics of the entire system.

Demonstrating the importance of pluralistic vertical integration in the Leabra theory, this large-scale specialisation of the cognitive architecture is suggested by basic neural mechanisms. For example, a single neural network cannot both learn general statistical regularities about the environment and quickly learn arbitrary new information such as new facts or names of people (McClelland, McNaughton, and O'Reilly 1995; O'Reilly and Rudy 2001; O'Reilly and Norman 2002). Specifically, rapid learning of arbitrary new information requires sparse, pattern-separated representations and a fast learning rate, whereas statistical learning requires a slow learning rate and overlapping distributed representations. These properties correspond nicely with known biological properties of the hippocampus and neocortex, respectively. A number of empirical studies, specifically motivated by Leabra computational modelling work, have tested and confirmed these and other more detailed properties (e.g. Maviel, Durkin, Menzaghi, and Bontempi 2004; Bakker, Kirwan, Miller, and Star 2008).
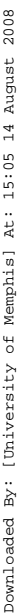
Similar reasoning applies to understanding the specialised properties of the frontal cortex, particularly the prefrontal cortex, relative to the posterior neocortex and hippocampal systems. The tradeoff in this case involves specialisations required for maintaining information in an active state (i.e. maintained neural firing, supported by the frontal cortex) relative to those required for performing semantic associations and other forms of inferential reasoning (supported by the posterior cortex). The prefrontal cortex system also requires an adaptive gating mechanism (Braver and Cohen 2000; O'Reilly and Frank 2006), to be able to rapidly update certain new information, such as a new subgoal, while simultaneously maintaining other information that remains relevant, such as a super-ordinate goal. The basal ganglia have the right neural properties to provide this function (Frank, Loughry and O'Reilly 2001).

The Leabra framework has been and continues to be applied to modelling a wide range of cognitive phenomena in perception, attention, learning and memory, language, and higher-level cognition, thereby testing and validating the synthesis of its core elements.

5. Theoretical convergence

When the ACT-R and Leabra research teams began working together in 2006, they came to a startling realisation: the two theories, despite their origins in virtually opposite paradigms (the symbolic and connectionist traditions, respectively) and widely different levels of abstraction, were remarkably similar in their view of the overall architecture of the brain. Furthermore, they discovered that the underlying subsymbolic mechanisms in ACT-R have conceptual and even mathematical similarity to the behaviour of emergent representations in Leabra. Finally, they recognised that each architecture reflects an explicit commitment to theoretical pluralism, both vertically (coarse *versus* granular) and horizontally (e.g. replaceable modules in ACT-R, multiple learning mechanisms and strategies in Leabra).

At the level of large-scale systems, the theoretical agreement is evident in Figure 3. Both architectures reflect a central role for the basal ganglia in receiving converging input from a wide range of cortical processing areas, which then drives the performance of

Figure 3. Overlap of the ACT-R and Leabra architectural organisation.

specific motor or cognitive actions. Anatomically, the basal ganglia send output primarily to the frontal cortex, which is associated with active maintenance of task relevant information in Leabra, and with the homologous buffers of ACT-R. Similarly, both architectures highlight the importance of the declarative/episodic memory system supported by the hippocampus and related anatomical structures. Finally, both adopt specialised sensory and motor processing pathways that have been well characterised in posterior cortex. In the Leabra architecture, the processing differences among these systems are supported by distinct neural specialisations in the basal ganglia, hippocampus and cortex, while in ACT-R they are supported by distinct modules, representational structures and processing and learning mechanisms.

At the cognitive level, a division into procedural and declarative components is shared by both frameworks. This distinction and dissociation has clear cognitive validity: people can possess abstract declarative knowledge of how to do something, yet be procedurally incapable of doing so (e.g. driving a car or playing golf), and vice-versa (e.g. touch typists often cannot recall where the keys are located).

Although dissociable, the procedural and declarative systems interact intimately in any complex cognitive process. In ACT-R, the firing of productions is driven by the active contents of the declarative and other information buffers, and the result of production firing is the updating of these buffers and the creation of new declarative memory chunks. In Leabra, the basal ganglia procedural system is tightly linked with the prefrontal cortex, which maintains task-relevant information in an active state over time. One of the primary functions of the basal ganglia in the brain is to drive the updating of these prefrontal active memory states. These prefrontal areas then influence activation states throughout the rest of the cortex via strong top-down excitatory projections. Each area of posterior cortex has an associated prefrontal area, with which it has strong bidirectional excitatory connectivity. Thus, we associate the buffers of ACT-R with these prefrontal representations of corresponding posterior cortical areas. While some buffers have been associated with activation of posterior cortical areas as well as prefrontal areas (e.g. Anderson 2007),

those attempts do not specifically discriminate between the activation of buffers and the modules with which they are associated.

Reinforcement learning is a central behavioural element of human and animal procedural cognition. The basal ganglia system in Leabra is strongly modulated by a model of dopamine, which signals reward and punishment information. Positive reward reinforces associated procedural actions, while negative feedback reduces the likelihood of producing associated actions. A similar, more abstract form of reinforcement learning is present in the ACT-R procedural system, where the integrated history of past success and time-cost are the major determinants in selecting which production will fire. Stochastic noise in quantities learned by reinforcement learning that control procedural selection plays an important role in both frameworks in modulating maximising tendencies, and suggests similar solutions to issues of exploration vs. exploitation.

The implementation of declarative memory in the two architectures stands in contrast to that of reinforcement learning. The neural properties of the hippocampus have been shown in the Leabra framework to be critical for the rapid learning of new arbitrary information without interfering with existing knowledge. The declarative system in ACT-R integrates both of these properties: new chunks of knowledge, encoded as combinations of existing chunks, can be rapidly formed and accessed unambiguously; chunks that are used more frequently over time gain higher levels of activation and correspond to more expert knowledge; similarities can be defined between symbolic chunks to drive semantic generalisation to related situations.

Both architectures also make use of associative learning mechanisms to modulate the strength of representations in declarative memory. Subsymbolic declarative quantities in ACT-R are learned according to Bayesian statistical algorithms, while new declarative representations in Leabra are learned using a combination of error-driven and Hebbian learning. Such learning mechanisms are based on the history of activation of the information stored in declarative memory, but, critically, not on the success, failure or costs of a particular action taken using that memory, as in procedural reinforcement learning. In terms of processing information already stored in declarative memory, the concept of spreading activation is critical to both architectures. In ACT-R, activation spreads among declarative chunks in proportion to their associative strength and similarities between chunks determine the degree of match. In Leabra, a similar activation spreading dynamic emerges, in that coarse-coded distributed representations in posterior cortical areas cause associated representations to overlap and share activation states.

The two architectures, however, reached different solutions regarding the existence of complementary declarative systems. While Leabra had to assume separate systems with distinct properties reflecting those of the hippocampus and posterior cortex because no neurally plausible learning rules could produce the properties of both systems, ACT-R was able to adopt a unified approach to declarative memory. The ability to create arbitrary combinations of existing symbolic chunks into new structures provides the rapid learning typical of the hippocampus, while the slow accumulation of activation and strengths of association reflects the slow, statistical learning of posterior cortical areas. More specifically, it was ACT-R's commitment to an integrated hybrid symbolic-subsymbolic approach that enabled it to unify separate areas into a single one. The result is that commitments to different mechanistic levels allows on the one hand for an understanding of the function of separate subsystems, but on the other for the unification of their underlying functions. Each level of the system thus makes distinct contributions to our scientific understanding.

In summary, the ACT-R and Leabra frameworks overlap to a considerable degree in their basic claims about the nature of the cognitive and neural architecture, despite having been developed from very different perspectives. Broadly speaking, they are mutually coherent theories of the same unity of behaviour, at different levels of description.

6. SAL

Like all theories, ACT-R and Leabra are incomplete. While ACT-R utilises subsymbolic mechanisms and can interact with modules that are not symbolic, ultimately its inputs and representations must translate to pure symbols. Leabra argues for a tripartite large-scale architecture, but this architecture has not yet been implemented in an integrated large-scale simulation. Neither of the architectures provides much detail regarding the crucial issue of how symbolic representations, the feature of human cognition that makes it unique among animals, arise organically in the mind and brain, although some initial work on this issue has begun (Rougier, Noelle, Braver, Cohen, and O'Reilly 2005).

Because a philosophy of pluralism is not only inherent in the theories, but also a part of each group's working style, there seemed to be promise in collaboration. Thus, the SAL architecture was born. SAL is an attempt to integrate and synthesise the Leabra theory of neural function, network behaviour and representation, and tripartite architecture with the ACT-R theory of symbolic and subsymbolic decision-making, representational activation and organisation, and modular architectural organisation. It also is worth pointing out that in the combined SAL architecture, most major machine learning techniques are represented, and grounded in forms that are motivated and informed by human psychology and biology.

For its initial effort, the SAL team built a demonstration model representing a preliminary synthesis of the two architectures (Figure 4). The model performs
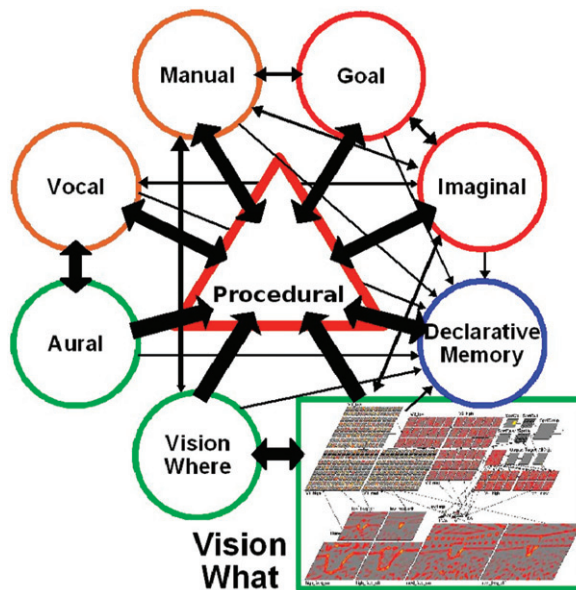

Figure 4. An initial SAL implementation.

a straightforward navigation and search task. The integration is of the simplest form, whereby the visual module in an existing ACT-R model of navigation is replaced with a Leabra visual object recognition model, which is capable of processing raw bitmap images in a way that the ACT-R visual module cannot. Similarly, extant Leabra models are not capable of organising problem solving behaviour over a period of several minutes, as the ACT-R model does in searching for the target object in a complex environment. Thus, this hybrid SAL model represents a new level of functionality that goes beyond the capabilities of its constituent architectures.

In the demonstration, the SAL agent is embodied within an Unreal Tournament simulation environment. It is familiar with the environment, in that it has access to navigation points and object location points in symbolic form. This simplification was adopted for tractability and without much loss of generality, as Best and Lebiere (2006) demonstrated that one ACT-R model can navigate either virtual or real worlds (controlling a virtual avatar or a robotic platform, respectively) using actual sensors without requiring pre-arranged navigation points. Further, its Leabra-based vision module has been trained to perceptually identify the possible object categories from bitmaps under a variety of viewing angles and distances. An operator instructs SAL to find the desired target via a typed command ('find armor'). SAL then navigates the rooms; views and perceptually identifies each object; and, when it recognises the desired target, navigates to it and picks it up (Figure 5).

This straightforward modular integration illustrated that significant behavioural benefits can arise from the synthesis. The obvious value is that the SAL model can accomplish a task that neither architecture could perform alone: ACT-R because its visual module cannot recognise objects from bitmaps, and Leabra because it has not demonstrated control properties that allow it to navigate complex spatial environments.
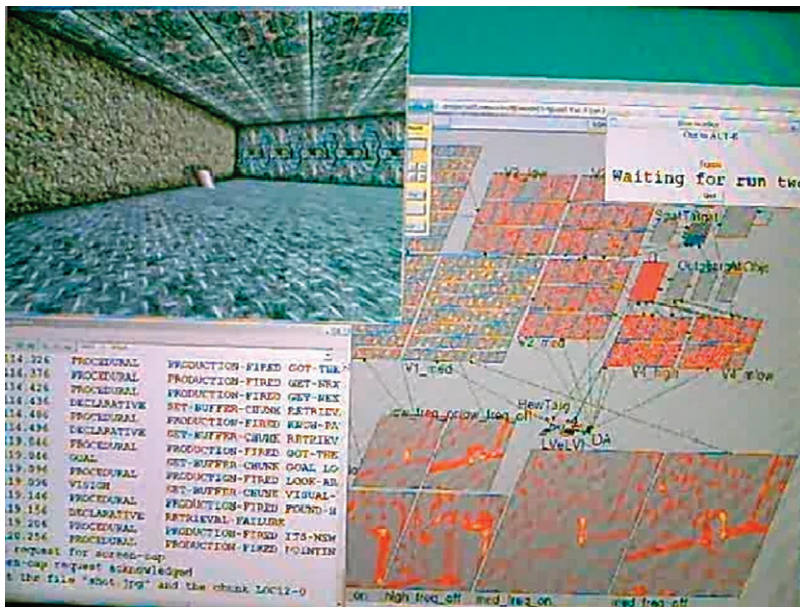

Figure 5. SAL operating in an Unreal Tournament environment.

However, the model also sheds light on some basic theoretical conundrums, such as the symbol-grounding problem. The SAL model demonstrates how the architecture could train the Leabra visual network to associate images of an object to its symbolic representation, and then use that capability to robustly recognise objects in the environment, extract their symbolic identity, and use that information to control complex behaviour. While, again for the sake of simplicity, that process was managed by the modelers in this case, it would be relatively straightforward to combine it with ACT-R's demonstrated ability to learn from instructions (e.g. Fu et al. 2004). This would result in a model that could be shown pictures of objects and told their names, and given arbitrary instructions involving those objects would be able to interpret and execute them.

Another interesting issue is the degree to which top-down control of the visual system by higher-level cognition modulates the bottom-up processing of visual inputs. In the current model, the model picks an object, focuses the attention of the visual module on that object, then requests the visual module to recognise it and receives the result. A more natural and efficient organisation would be to prime the visual module with the identity of the object being searched, then to allow the visual module to select the object in the visual field that best matches that description. Integrating a biologically plausible visual processing module with high-level control modules allows for the systematic investigation of those issues in tasks and environments of much greater complexity than those typically used in purely experimental investigations.

Taatgen, Juvina, Herd, Jilk and Martens (2007) showed that a deeper integration can produce more subtle behavioural features of human performance. Specifically, their model of attentional blink connected the temporally varying activation states of internal Leabra visual representations with an ACT-R model of control. In addition to demonstrating the primary phenomenon of attentional blink, it also showed subtleties such as reduced control leading to less blink, and reversed report of lag-one trials, neither of which were captured in prior symbolic or neural models. This empirical match suggests that the coherence and convergence of SAL, as described above, is not merely theoretical.

7. Continuing research

Future work on SAL will proceed along two different tracks. The first track can be likened to comparative anatomy, in that the structural and emergent features of each architecture will be mapped to those of the other, driving deeper understanding of both. This mapping will be informed by empirical biological and behavioural data. The second track emphasises achievement of superior functional capabilities through tighter, principled integration of the two simulation systems.

7.1. *Mapping track*

The theoretical issue at the heart of the integration between ACT-R and Leabra is how we think of the mapping between their structures and representations, most specifically between the ACT-R buffers and production rules, and the Leabra model of working memory and cognitive control in the prefrontal cortex and basal ganglia. Thus the effort will focus on the issue of neural realisations of the basic production-rule cycle in ACT-R (Figure 6), in which the contents of buffers are consulted, an action is selected, and the buffer contents are updated. The envisioned resulting dynamic is one of gradual
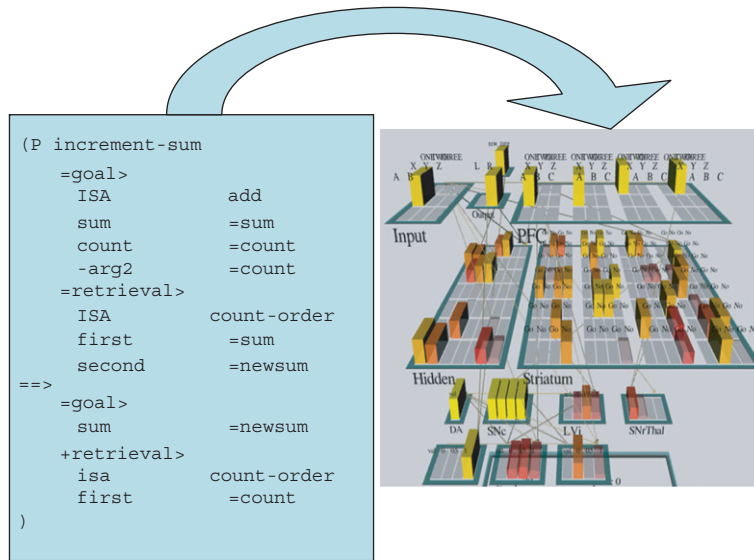
```
(P increment-sum
   =goal>
     ISA           add
     sum           =sum
     count         =count
     -arg2         =count
   =retrieval>
     ISA           count-order
     first         =sum
     second        =newsum
==>
   =goal>
     sum           =newsum
   +retrieval>
     isa           count-order
     first         =count
)
```

Figure 6. Mapping track – neural realisation of the production rule cycle.

convergence between two different and currently incommensurable levels of description, with each level gaining functionality and fidelity, while receiving guidance toward helping the other achieve its goals. Specifically, ACT-R will provide a control framework to guide the evolution of Leabra, while Leabra will provide a representation framework to constrain the evolution of ACT-R. Further, the final outcome of this process will result in a comprehensive account of cognition that ties behaviour at the organism level to mechanism at the cellular and sub-cellular level without explanatory gaps.

In particular, we anticipate that this research track will help to address the following open questions in ACT-R and Leabra:

- In ACT-R, how can 'partial matching' of production rules that 'softens' traditional production conditions operate without producing degenerate behaviour?
- How can the learning of utility parameters that control production selection be grounded in plausible assumptions about the nature of feedback available?
- How can context influence production matching and selection beyond the explicit specification of precise conditions?
- In Leabra, how can new procedural learning piggyback on prior learning, rather than using trial and error for each new behaviour?
- How can verbal instructions be integrated and processed?
- How does the system manage the choice between exploration and exploitation, and more generally perform strategic search through the space of behavioural actions and policies?

The neural realisation approach has informed the ACT-R architecture throughout its evolution, from an early attempt at implementing ACT-R in neural network constructs (Lebiere and Anderson 1993) to a current attempt at mapping the details of the production rule cycle onto the anatomy of the basal ganglia (Stocco, Lebiere, and Anderson 2008).

Developing neural realisations of higher-level cognitive phenomena has also been at the heart of the Leabra research program (O'Reilly and Munakata 2000), and this effort for the first time offers an intermediate level of representation that more tightly constrains these neural realisations. This history provides a great deal of confidence that important new insights will arise from the effort.

7.2. *Integration track*

As mentioned, the second future research track will emphasise the development of superior functional capabilities through a principled integration of the two systems. A critical feature of robust systems is that they have multiple different methods of solving problems. Those different methods have complementary strengths and weaknesses that prevent catastrophic failures (because when one method fails, another can take over) and boost the overall performance of the system (by both synthesising the results of the different methods and allowing some methods to learn from others). Human intelligence in particular gains considerable robustness by having both subsymbolic and symbolic capabilities. O'Reilly (2006) recently characterised these two capabilities in terms of analog and digital computational properties that emerge from computational models based on cognitive neuroscience data.

The division between symbolic and subsymbolic (i.e. distributed) levels corresponds largely to the distinction between control and representation subsystems, respectively. A combination of the two is essential in implementing a broadly effective system, but as one would expect, and as some past efforts have shown, the specific manner in which they are combined is a critical determinant as to whether the potential functional benefits of the combination are realised. Neurally realised subsymbolic systems provide a way to achieve the following properties in processing and representing large amounts of data:

- Speed: distributed connectionist systems can process information in a few steps by exploiting neural parallelism.
- Capacity: massive parallelism at the cortical level enables the processing of large amounts of data in perceptual and memory areas.
- Robustness: unlike symbolic systems, distributed representations generalise naturally to new situations and their performance degrades gracefully in the face of erroneous, imprecise or unexpected information or damage to the system.

Conversely, advantages of symbolic systems for controlling the operations of the system include:

- Tractability: Enforcing sequentiality sacrifices speed to enable tractable control over the flow of execution and reduce combinatorial complexity.
- Inspectability: Sequential control steps enable explicit memory of past processing and the metacognitive introspection essential to avoiding local solution minima (i.e. impasses).
- Efficiency: learning sequential control is considerably more efficient for symbolic systems than it is for subsymbolic systems.

Past attempts to develop symbolic/subsymbolic hybrid systems (see, e.g. Wermter and Sun 2000 for a review) have typically not reflected this division of labour. Instead of applying symbolic techniques to control functions and subsymbolic techniques to

representation functions, they often use both techniques for both functions. This dilutes rather than exploits the benefits of each type of technique, resulting in systems that struggle in both control and representation of the information required for robust cognition.

In SAL there is a simple and obvious mapping of ACT-R to the control of the communication paths among brain regions, and of Leabra to the subsymbolic representation and computations performed within those regions; the research will emphasise their strengths in these domains. However, we plan to go beyond a simple modular hybridisation scheme toward a deeper synthesis, where elements based on ACT-R and Leabra principles interact in a more tightly coupled and biologically inspired manner. In brief, we associate the active buffers, procedural production system, and symbolic representations from ACT-R with the prefrontal cortex and basal ganglia, while the graded distributed representations and powerful learning mechanisms from Leabra are associated with the posterior cortex. Given these mappings and specialisations, in the SAL framework the interface between symbolic and subsymbolic elements occurs in the bidirectional interactions between the ACT-R elements associated with prefrontal cortex, and the Leabra elements associated with posterior cortex. Thus the tight integration between ACT-R and Leabra occurs not *between* modules, as in our demonstration system, but *within* modules; and in the process separating the larger subsymbolic module from its symbolic buffer interface to the procedural system.

This direction differs substantially from the first track, in which the focus is on developing neural realisations of working memory and cognitive control. Here, ACT-R primarily handles cognitive control and Leabra handles representation, and the emphasis is on the bidirectional interface between the two (Figure 7).

There are important technical challenges to bridging the analog/digital gap at the interface of the posterior and prefrontal elements of this model, but we believe that these are the very same challenges that the human brain faces in order to exhibit both symbolic and subsymbolic abilities. Thus, a major focus of the research will be on the specific issue of understanding bidirectional interactions between prefrontal and posterior representations, especially how learning new distributed representations in posterior cortex can drive the development of new prefrontal symbolic representations, and how the task demands represented in prefrontal cortex provide top-down influences that shape the learning and processing in posterior cortex. The human system clearly gains considerable power and
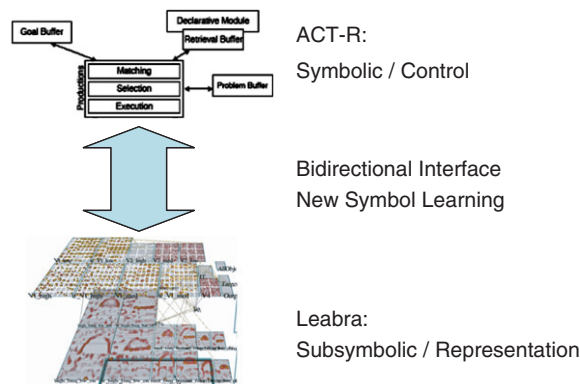

Figure 7. Integration track – bidirectionally connected processes.

robustness from a bidirectional synergy between these systems, and we think that capturing this dynamic is the key to making SAL more than the sum of its analog and digital parts. SAL has an advantage in this area over past attempts at hybridisation, in that Leabra naturally incorporates bidirectional learning and activation dynamics, and ACT-R is the only symbolic architecture that incorporates subsymbolic computations to perform both bottom-up statistical learning and top-down biasing of information-processing functions such as memory retrieval.

Beyond these two primary research thrusts, SAL principles are currently being applied in an existing robotics project where ACT-R is the control system and a variety of machine learning and algorithmic perception approaches are used. A Leabra object recognition model is being introduced to work in conjunction with region-of-interest detection algorithms to identify objects in the scene. Spatial attention has proved challenging to model successfully, and in addition to the functional benefits, we hope that this collaboration will illustrate constraints on how spatial attention must operate.

A pluralistic approach is also being used in a project to combine neural and algorithmic approaches to perception in simulated environments as a replaceable 'front-end' to symbolic cognitive architectures. Leabra and ACT-R will serve as the reference implementation, but the project aims to treat them as modular and replaceable components, as long as such components fit within the needs and constraints of cognition.

8. Discussion

We have provided considerable detail of the theoretical overlap between ACT-R and Leabra, the previous and planned approaches to their integration in SAL, and the theoretical questions we hope to answer through this effort. This detail shows that the SAL architecture is *explicitly* pluralistic, not merely in that its constituent architectures exhibit vertical, modular and mechanistic pluralism, or in the simple fact that it is a hybrid, but rather that this hybrid maps to dissociable systems in the human brain, and aims to integrate those systems in a manner similar to the brain. In the words of Newell (1990):

> A single system (mind) produces all aspects of behaviour. It is one mind that minds them all. Even if the mind has parts, modules, components or whatever, they all mesh together to produce behaviour. Any bit of behaviour has causal tendrils that extend back through large parts of the total cognitive system before grounding in the environmental situation of some earlier times. If a theory covers only one part or component, it flirts with trouble from the start. It goes without saying that there are dissociations, independencies, impenetrabilities and modularities. These all help to break the web of each bit of behaviour being shaped by an unlimited set of antecedents. So they are important to understand and help to make that theory simple enough to use. But they do not remove the necessity of a theory that provides the total picture and explains the role of the parts and why they exist.

On the surface, ACT-R and Leabra are incommensurable: one operates on discrete chunks and production rules, while the other is based on simulated neurons and their interconnections. While we hope to map the theories either mathematically or in simulated form, the incommensurable categories at the various levels of description will remain necessary to explain the full range of phenomena (Rohrlich 1988).

The SAL architecture intends to explain a broader range of phenomena, and simulate a larger scope of functionality, than can either of the component architectures alone.

However, even if it proves successful, we will continue to view SAL as an epistemologically attractive description of certain aspects of cognition, not as an ontologically privileged theory. Other types of theories can constrain or explain SAL in further directions to continue to broaden its scope or improve its performance. For example:

- The SOAR cognitive architecture (Newell 1990) has explained aspects of metacognition, including theory of mind and self-assessment. Given that these aspects are central to human social function, how might they constrain the representations in SAL? Are new architectural features required in ACT-R to model metacognition? Do mirror neurons have specialised properties or do they emerge out of the representations learned in cognition?
- Machine learning algorithms, such as Bayesian approaches and reinforcement learning, characterise mathematically optimal solutions to problems. Homologous features of SAL should approach such optimality, explain why such features are not optimal in a broader cognitive context, or illustrate that the human system is suboptimal due to unrelated biological factors, in effect extending the rational analysis (Anderson 1990) to include factors at multiple levels of abstraction.
- Can artificial cognitive systems shortcut traditional human learning? Is it possible to 'load' a corpus of *a priori* knowledge, such as the Cyc ontology (Lenat and Guha 1990), without an elaborate educational process and without semantic loss? If so, can this extend down to motor and perceptual learning or does it apply only to logical categories and relationships?
- New molecules and pathways that are required for long-term potentiation of synapses, or of *in vivo* learning, are constantly being discovered. Which elements of this neurochemical maelstrom produce important computational features? Are there simplified mathematical descriptions of these features or is it necessary to simulate them?

Finally, there have been explicit benefits to the pluralistic collaboration itself. One such benefit is characterised by the aphorism 'you don't learn it until you teach it'. Because the collaborators must have a fairly thorough understanding of the component architectures, we have recognised areas where there theories could use further elaboration or explanation. In attempting to connect the two architectures functionally, it became clear that the question of how symbols arise is unanswered in both. The very pragmatic process of writing joint proposals provides insight into the core questions that collaborators emphasise and how they sell this to funding organisations – thus allowing one to improve the contextual element of separate proposals.

We have found a pluralistic approach to cognitive science to be highly fruitful both within the scope of individual theoretical constructions and in the context of collaboration. In the spirit of pluralism, however, we recognise that others may have a different view.

Authors Disclosure of Interest Statement

related to SAL have been developed to date, but the Authors write SAL proposals in which eCortex, Inc., is a contributor and thus a financial beneficiary.

References

Anderson, J.R. (1976), *Language, Memory, and Thought*, Hillsdale, NJ: Erlbaum.
———. (1983), *The Architecture of Cognition*, Cambridge, MA: Harvard University Press.
———. (1990), *The Adaptive Character of Thought*, Hillsdale, NJ: Erlbaum.
———. (1993), *Rules of the Mind*, Hillsdale, NJ: Erlbaum.
———. (2002), 'Spanning Seven Orders of Magnitude: A Challenge for Cognitive Modeling', *Cognitive Science*, 26, 85–112.
———. (2007), *How can the Human Mind Occur in the Physical Universe?*, New York, NY: Oxford University Press.
Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., and Qin, Y. (2004), 'An Integrated Theory of Mind', *Psychological Review*, 111(4), 1036–1060.
Anderson, J.R., and Bower, G.H. (1973), *Human Associative Memory*, Washington: Winston and Sons.
Anderson, J.R., and Gluck, K. (2001), 'What Role do Cognitive Architectures Play in Intelligent Tutoring Systems?' in *Cognition & Instruction: Twenty-five years of progress*, eds. D. Klahr and S.M. Carver, Mahwah, NJ: Erlbaum, pp. 227–262.
Anderson, J.R., and Lebiere, C. (1998), *The Atomic Components of Thought*, Mahwah, NJ: Erlbaum.
Bakker, A., Kirwan, C.B., Miller, M., and Star, C.E.L. (2008), 'Pattern Separation in the Human Hippocampal CA3 and Dentate Gyrus', *Science*, 319, 1640–1642.
Best, B.J., and Lebiere, C. (2006), 'Cognitive Agents Interacting in Real and Virtual Worlds', in *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, ed. R. Sun, New York, NY: Cambridge University Press, pp. 186–218.
Braver, T.S., and Cohen, J.C. (2000), 'On the Control of Control: The Role of Dopamine in Regulating Prefrontal Function and Working Memory', in *Control of Cognitive Processes: Attention and Performance*, eds. S. Monsell and J. Driver, Cambridge, MA: MIT Press, pp. 713–737.
Byrne, M.D., and Kirlik, A. (2005), 'Using Computational Cognitive Modeling to Diagnose Possible Sources of Aviation Error', *International Journal of Aviation Psychology*, 15, 135–155.
Feyerabend, P. (1993), *Against Method* (3rd ed.), New York: Verso.
Feynman, R. (1965), *The Character of Physical Law*, Cambridge: MIT Press. (1994 reprint edition).
Frank, M.J., Loughry, B., and O'Reilly, R.C. (2001), 'Interactions Between the Frontal Cortex and Basal Ganglia in Working Memory: A Computational Model', *Cognitive, Affective, and Behavioral Neuroscience*, 1, 137–160.
Freeman, K. (1947), *Ancilla to the Pre-Socratic Philosophers*, Cambridge: Harvard University Press. (1983 reprint edition).
Fu, W.-T., Bothell, D., Douglass, S., Haimson, C., Sohn, M.-H., and Anderson, J.R. (2004), 'Learning from Real-Time Over-the-Shoulder Instructions in a Dynamic Task', *Proceedings of the 6th International Conference on Cognitive Modeling*, Pittsburgh, PA: Carnegie Mellon University/University of Pittsburgh, pp. 100–105.
Jilk, D.J., Cer, D.M., and O'Reilly, R.C. (2003), Learning Rules Generated by a Biophysical Model of Synaptic Plasticity. *Computational Neuroscience Conference*.
Kuhn, T.S. (1970), *The Structure of Scientific Revolutions* (2nd ed.), Chicago: The University of Chicago Press.
Lebiere, C., and Anderson, J.R. (1993), 'A Connectionist Implementation of the ACT-R Production System', in *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 635–640.

Lebiere, C., Anderson, J.R., and Reder, L.M. (1994), 'Error Modeling in the ACT-R Production System', in *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum, pp. 555–559.

Lenat, D., and R. Guha, V. (1990), *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*, Reading, MA: Addison-Wesley.

Lisman, J., Lichtman, J., and Sanes, J. (2003), 'LTP: Perils and Progress', *Nature Reviews Neuroscience*, 4, 926–929.

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

Maviel, T., Durkin, T.P., Menzaghi, F., and Bontempi, B. (2004), 'Sites of Neocortical Reorganization Critical for Remote Spatial Memory', *Science*, 305, 96–99.

McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995), 'Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory', *Psychological Review*, 102, 419–457.

Meyer, D.E., and Kieras, D.E. (1997), 'A Computational Theory of Executive Control Processes and Human Multiple-Task Performance: Part 1. Basic Mechanisms', *Psychological Review*, 104, 3–65.

Newell, A. (1972), 'A Theoretical Exploration of Mechanisms for Coding the Stimulus', in *Coding Processes in Human Memory*, eds. A.W. Melton and E. Martin, Washington, DC: Winston, pp. 373–434.

———. (1973a), 'Production Systems: Models of Control Structures', in *Visual Information Processing*, ed. W.G. Chase, New York, NY: Academic Press, pp. 463–525.

———. (1973b), 'You Can't Play 20 Questions with Nature and Win', in *Visual Information Processing*, ed. W.G. Chase, New York, NY: Academic Press, pp. 283–308.

———. (1990), *Unified Theories of Cognition*, Cambridge, Massachusetts: Harvard University Press.

O'Reilly, R.C. (1996), 'Biologically Plausible Error-Driven Learning Using Local Activation Differences: The Generalized Recirculation Algorithm', *Neural Computation*, 8, 895–938.

———. (1998), 'Six Principles for Biologically-Based Computational Models of Cortical Cognition', *Trends in Cognitive Sciences*, 2, 455–462.

———. (2001), 'Generalization in Interactive Networks: The Benefits of Inhibitory Competition and Hebbian Learning', *Neural Computation*, 13, 1199–1242.

———. (2006), 'Biologically Based Computational Models of High-Level Cognition', *Science*, 314, 91–94.

O'Reilly, R.C., and Frank, M.J. (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., Frank, M.J., Hazy, T.E., and Watz, B. (2007), 'PVLV: The Primary Value and Learned Value Pavlovian Learning Algorithm', *Behavioral Neuroscience*, 121, 31–49.

O'Reilly, R.C., and Munakata, Y. (2000), *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, Cambridge, MA: MIT Press.

O'Reilly, R.C., and Norman, K.A. (2002), 'Hippocampal and Neocortical Contributions to Memory: Advances in the Complementary Learning Systems Framework', *Trends in Cognitive Sciences*, 6, 505–510.

O'Reilly, R.C., and Rudy, J.W. (2001), 'Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function', *Psychological Review*, 108, 311–345.

Rohrlich, F. (1988), 'Pluralistic Ontology and Theory Reduction in the Physical Sciences', *The British Journal for the Philosophy of Science*, 39, 295–312.

Rohrlich, F., and Hardin, C.L. (1983), 'Established Theories', *Philosophy of Science*, 50, 603–17.

Rougier, N.P., Noelle, D., Braver, T.S., Cohen, J.D., and O'Reilly, R.C. (2005), 'Prefrontal Cortex and the Flexibility of cognitive control: Rules Without Symbols', *Proceedings of the National Academy of Sciences*, 102, 7338–7343.

Salvucci, D.D. (2006), 'Modeling Driver Behavior in a Cognitive Architecture', *Human Factors*, 48, 362–380.

Stocco, A., Lebiere, C., and Anderson, J.R. (2008), 'Procedural Learning and Sequential Control of Behavior in a Neural Network Model of the Basal Ganglia'. To be presented at 2008 Computational Neuroscience Meeting.

Taatgen, N.A., Juvina, I., Herd, S., Jilk, D., and Martens, S. (2007), 'Attentional Blink: An Internal Traffic Jam?' in *Proceedings of the 8th International Conference on Cognitive Modeling*. Ann Arbor, Michigan, USA.

Wermter, S., and Sun, R. (Eds.) (2000). *Hybrid Neural Systems*. Heidelberg: Springer Verlag.