International Journal of Machine Consciousness
Vol. 2, No. 2 (2010) 171–192

DOI: 10.1142/S1793843010000552

THE B-I-C-A OF BIOLOGICALLY INSPIRED COGNITIVE ARCHITECTURES*

ANDREA STOCCO

Institute for Learning and Brain Sciences
University of Washington, Seattle, WA 98195

CHRISTIAN LEBIERE

Department of Psychology, Carnegie Mellon University,
Baker Hall 345-A, 5000 Forbes Avenue
Pittsburgh, PA 15213, USA
cl@cmu.edu

ALEXEI V. SAMSONOVICH

Krasnow Institute for Advanced Study,
George Mason University, 4400 University Drive MS 2A1,
Fairfax VA 22030-4441, USA
asamsono@gmu.edu

Recent years have seen a gradual convergence of seemingly distant research fields over a single goal: understanding and replicating biological intelligence in artifacts. This work presents a general overview on the origin, the state-of-the-art, scientific challenges and the future of Biologically Inspired Cognitive Architecture (BICA) research. Our perspective decomposes the field into the four principal semantic components associated with the BICA challenge that together call for an integration of efforts of researchers across disciplines. Areas and directions of study where new integrated efforts will be primarily needed are summarized.

Keywords: Cognitive architectures; brain reverse-engineering; biological constraints; human-level intelligence.

1. The BICA Doctrine

A *cognitive architecture* is a computational model or framework used for the design of intelligent agents. Biologically Inspired Cognitive Architectures, or BICA, emerge at the intersection of computational and brain sciences, as the new integrative paradigm in non-von-Neumann computing, unifying other popular paradigms in artificial

*A manifesto of the emergent BICA community represented at the BICA 2010 conference: Arlington, Virginia, USA, 12–14 November 2010 (http://bicasymposium.com, http://roboticslab.dinfo.unipa.it/bica2010/).

intelligence (AI) including connectionism, evolutionary computation, nonlinear dynamical systems, Bayesian networks, production systems, and many others. Why and what makes this new paradigm different and better? The answer can be found in the foundational ideas of BICA. One of these ideas is completeness: in contrast with previous limited attempts, the BICA ambition is to develop a complete biologically inspired agent capable of intelligent behavior in realistic environments. This is complemented by the idea of biological inspiration, which has two aspects. One is the borrowing of information processing and learning principles from biology, where they are known to yield the unparalleled robustness, flexibility and evolvability of natural intelligent systems. Another aspect of biological inspiration is the desire to achieve cognition in artifacts that is similar to human cognition, so that the artifacts would integrate better into the human society. An artifact of this kind will be human-compatible in the strong sense: e.g., understandable by humans from its behavior, trustworthy, communicative, sensible, friendly, capable of human-like emotions (at least in its apparent behavior), reasoning and learning.

Here the last detail, if taken seriously, becomes the most important foundational idea of the BICA doctrine. It can be called *the principle of cognitive growth.* This idea simply turns upside down the original paradigm of AI, which was to develop, implement and integrate strong, powerful, sophisticated, often optimal capabilities. While it is true that this process can result in synergy, it did not result in a major leap during the last 50 years of AI. In contrast, the BICA challenge is to find a design as weak and simple as possible, limiting it to the bare *critical mass* of intelligent capabilities that enable human-level learning. That is, the critical mass from which the self-sustained cognitive growth can take off and go on up to the adult human level of general intelligence.

But, does this notion of "critical mass" make sense in the context of human-level general intelligence? Examples from biology suggest a positive answer to this question. Neuroanatomically, the difference between the ape brain and the human brain essentially amounts to minor differences in sizes of brain structures [e.g., Semendeferi and Damasio, 2000]. Even the rat brain is used as a model of the human brain in neurophysiological studies of higher cognitive functions [e.g., Babb and Crystal, 2006]. It has been reported that certain birds, for example, seem capable of "theory of mind" [Dally *et al.*, 2010], learn to make and use tools [Hunt, 1996], develop abstract concepts and learn to communicate in English [Pepperberg and Gordon, 2005]. Yet, a huge gap separates humans from other animals in the general ability to learn. No non-human animal (and no computer today) can be trained like a student, e.g., to understand linear algebra and pass an exam in it. Therefore, it appears that a small step in the functional organization of the brain can make a huge difference. The present state-of-the-art in BICA research suggests that we may be close to the point where we can make this step that would allow us to solve the BICA challenge. *The essence of the BICA challenge is to replicate the human general ability to learn in artifacts.*

Every approach to developing a human-level artificial intelligence has its major obstacle. The major problem of the BICA challenge understood as defined above appears to be in the identification of the critical mass. Once precisely identified, it can be built. Below we consider the four major forces and needs (the "pillars") that drive BICA research today. We symbolically associate them with the four letters of the acronym: "B", "I", "C", "A", and explain this association below.

Although distinct, these four elements are connected by many common threads, often involve the same problems, and frequently benefit each other. Our hope is that the four major forces, when put together, will allow us to solve the ultimate BICA challenge. Therefore, we call for a unification of efforts across research communities and fields toward an integrated approach that leverages progress across domains and builds on advances in related fields.

2. The Four Pillars of BICA

In recent years, four major forces concurred to focus research on biologically inspired integrated intelligent systems. We briefly outline them below.

A first, bottom-up, driving factor has been the increased understanding of biological systems at the level of elements, which has resulted in the possibility of interpreting the brain's biological circuits with more formal computational abstractions than it was done in the early years of neuroscience, opening up bridges to higher-level AI research (the "reverse-engineering the brain" challenge, Albus [2010]).

The second driving force is the push for a unification across humanities, engineering and neurosciences, a manifestation of which is the series of recent attempts to develop a science of the mind [Albus *et al.*, 2007] or a science of consciousness. The idea is that we need more than just a statistical model that allows us to make good experimental predictions. The fundamental human drive to learn and desire to survive and propagate leads us to think about conscious machines that one day will become an extension of our culture and ourselves. The brain is ultimately an information-processing device, and there seem to be no objective reason that forbids the replication of the same principles in a different substrate.

The third element is the capability of re-using, distributing, and integrating models. Recent years have seen the advent of many integrated models, capable of bridging the gap between a plausible structural simulation of neuronal circuits and high-level behavior in complex tasks. These developments have been instrumental in introducing biological constraints as a means to compare alternative theories. They also pose new challenges in terms of model distribution, comparison, re-use and data sharing via repositories. Meeting these challenges would provide for faster progress toward longer-lasting, robust models.

Finally, the fourth, top-down, force has been the increased research on cognitive and integrated architectures, which has achieved unexpected successes and has increasingly found inspiration and challenge in the comparison to real, biological systems. Those architectures attempt to provide an account of general intelligence,

rather than task-specific attempts that have come to dominate the field of artificial intelligence recently.

According to the above, the four main scientific views and associated schools of thought in BICA research can be listed as follows: (B) computational neuroscience, that tries to understand how the brain works in terms of neurophysiological mechanisms and neuroarchitectures; (I) human-mind-inspired artificial intelligence, aiming at anthropomorphic artificial minds that can be understood by humans intuitively, that can learn like humans, from humans and for human needs, and therefore can replace humans at work; (C) the challenge of cognitive modeling that pursues higher-level computational description of human cognition and behavior based on higher-level abstract models and cognitive architectures; and (A) architectures implemented in real artifacts that put the above ideas together. These four fundamental scientific approaches labeled by the letters of the acronym BICA are addressed in detail below.

3. "B" for Biology: The Bottom-Up Reverse-Engineering of the Brain

The human brain is the gold standard for comparing artificial intelligent systems. In the history of computer science, the brain has been the inspiration for the introduction of a number of computer science techniques, like perceptrons [Rosenblatt, 1958] and networked computers [Licklider, 1968]. However, limitations of these early ideas were soon understood [Minsky and Peipert, 1969].

Later, the connectionist revolution in the 1980s opened up the possibility of creating artificial neural networks capable of performing computationally interesting tasks [Hopfield, 1982; Rumelhart and McClelland, 1986]. The enthusiasm of the original connectionist revolution had been somewhat curbed by the understanding that very little was known at the time about important features of the biological circuits, and that the similarity between artificial and biological neural networks was more limited than originally estimated.

Brain research in recent years has shed light on a number of issues, including the way neurons encode and transmit information, the way information is encoded across large ensembles of neurons, and the way neuronal ensembles can be modeled to achieve large-scale simulations. Behind these advances is a story of bi-directional exchanges between AI and the brain sciences, which will be outlined in the sections below.

3.1. *From AI algorithms to brain signals*

The renewed connection between brain biology and artificial intelligence has also been fostered by the interpretation of certain neural signals in terms of well-understood algorithms that can be usefully adopted at a more symbolic level.

The most glaring example is perhaps reinforcement learning. The reinforcement learning problem is an optimal control problem where an agent has to learn the value of its available actions based on environment feedback [Sutton and Barto, 1998]. A particularly elegant and successful formulation was given in 1988 by Richard Sutton

[Sutton, 1988], in a paper where he discussed the possibility of estimating the actual value of an action by comparing two successive estimates of its value. The simple rationale behind this idea is that, while the actual value of an action is initially unknown, estimates become increasingly accurate with experience, and therefore the difference between successive estimates can be taken as a reliable error measure. Because of the nature of its estimates, this algorithm is known as Temporal Difference (TD) learning.

In a series of studies, Schultz and colleagues [Schultz *et al.*, 1993; Schultz, 1998] found that the response of dopamine neurons projecting to the basal ganglia matched very closely the changes in the error term in TD-learning. Subsequent studies and computational analysis have confirmed the similar nature of the two signals [Niv *et al.*, 2005].

The match between a universal biological mechanism and an AI algorithm, which was initially developed independently of neurophysiological considerations, is surprising in itself. Its importance, however, extends beyond this simple identification. The error term in the TD-learning algorithm presupposes a number of preliminary computations, including the estimates of the values of two consecutive states and the calculation of their differences. It is also possible to entertain a computational architecture that would make use of the calculations of the error term, most notably the so-called actor-critic architecture. Given the identification of the dopamine signal with the error term, it was possible to identify the functions of different parts of the circuit around the dopamine neurons by matching the required preliminary (or subsequent) computations with the neuron populations that lay upstream (or downstream) of the dopamine cells. This has led to a number of important clarifications regarding the nature of the basal ganglia circuit [e.g., Barto, 1995].

A second example is the nature of computations in the hippocampus. The hippocampus is a folding of the archicortex inside the temporal lobe cortex. Although in humans it is crucial for the formation of episodic memories [Scoville and Milner, 1957], its most basic function is perhaps the creation of spatial representations. In particular, studies in rats have identified two types of cells with remarkable properties in the hippocampal structure: grid cells in the medial entorhinal cortex [Hafting *et al.*, 2005; Fyhn *et al.*, 2007] that fire when the animal crosses the intersections of imaginary lines that divide the environment in a grid, and place cells in the hippocampus whose activity is tied to the animal being in a specific place [O'Keefe and D'Ostrovsky, 1971]. The nature of these two cell types has been debated, but their computations can be explained as intermediate results of path-integration algorithms [McNaughton *et al.*, 1996; Samsonovich and McNaughton, 1997; McNaughton *et al.*, 2006; Meyer and Kieras, 1997; Witter and Moser, 2006; Touretzky and Muller, 2006]. It has also been shown that place cells can be seen as the computational output of path-integrating over grid cell representation [Savelli and Knierim, 2010].

In summary, algorithmic solutions to AI problems have provided powerful means to interpret the activity of populations of neurons in the brain, and provided an initial starting point to speculate on the functions of larger circuits.

3.2. *Advances in technology*

The convergence in trying to understand and reproduce the biology of intelligence would not have been possible without integration of technological improvements in the brain sciences with advances in cognitive modeling. Many computational cognitive models of brain functions have been traditionally compared against behavioral data obtained by neuropsychological patients [e.g., Plaut, 2002]. A small number of models have been matched against data obtained from event-related potentials (ERP), that is, surface recordings of field potentials in the brain. The method of ERP found interesting applications to biological models of error correction [Yeung *et al.*, 2006] and word learning and hemispheric specialization [Mills *et al.*, 2005].

At the end of the 1980s, the introduction of a variety of new brain imaging techniques, including position-emission tomography (PET) and functional magnetic resonance imaging (fMRI), opened the possibility of non-invasive spatiotemporal mapping of neuronal activity involved in the execution of specified cognitive tasks [e.g., Posner and Raichle, 1994]. Initial studies were limited to the adoption of "block" designs, i.e., simple experimental manipulations that lack the capability of detecting interaction effects. This initial limitation was alleviated by the adoption, over the years, of a number of new statistical techniques for the analysis of fMRI data, and many software packages are now available that permit quite sophisticated data processing [e.g., Friston *et al.*, 2006].

The increased sophistication in data analysis also reflects an increased consciousness that computations of complex functions are carried out at a network level. If earlier studies were focused on identifying single regions by comparing activity patterns across different tasks, more recent experiments now permit the analysis of how a task is carried out in the functional network of brain structures, and the introduction of functional connectivity analysis has been given the opportunity to examine the communication between processing centers. A number of architectures have been proposed that describe the functional network dynamics [Anderson *et al.*, 2007; Just and Varma, 2007].

3.3. *The ultimate architecture: brain connectivity*

Functional connectivity ultimately relies on the existence of "structural" connectivity, i.e., physical bundles of axonal projections that connect distinct brain regions. The identification of these projections has traditionally been performed using various biochemical and genetic staining techniques that typically require *in vitro* analysis and therefore do not allow for temporal resolution.

The recent introduction of Diffusion Tensor Imaging [DTI; Le Bihan *et al.*, 2001] opens up the opportunity of tracking the structural connectivity between brain regions. DTI is based on the fact that physical bundles of fibers are oriented along particular directions in space, thus creating a preferential axis ("anisotropy") for molecules to diffuse. Spatial anisotropy can be analyzed, thus allowing the

reconstruction of structural connectivity. The Human Connectome Project [Sporns *et al.*, 2005] aims at reconstructing and validating the entire connectivity map of the human brain. Such a map would be invaluable in providing a detailed picture of the physical architecture of the human brain. In turn, a detailed physical architecture can provide a strong ground truth to evaluate competing biological models, and a strong foundation for reverse-engineering cognition.

3.4. *Other technologies*

In addition to the impressive growth of fMRI, other new technologies have permitted better insights into the brain. One of the limits of imaging techniques is that they are correlational in nature, and cannot permit causal inference. Suppose, for instance, that two areas are frequently co-activated by a set of tasks: There is no way to infer whether they are both needed to perform the task, or whether they perform the same computation independently and in parallel, or whether only one is needed and the activation of the second region is merely coincidental — due, for instance, to a common underlying input, or due to the unavoidable but spurious transmission of a signal along pathways that evolved for other reasons. There is no way to test these alternative hypotheses unless one interferes with the workings of one of the regions.

Fortunately, it is possible to perform something akin to "virtual" and reversible lesion of a brain region *in vivo* by using Transcranial Magnetic Stimulation [TMS; Hallett, 2000]. TMS involves the use of a powerful magnetic field to induce a short-lived, transient electric current in a particular area of the cortex. Inducing this current repeatedly results in a temporary down-regulation of the target neural population.

In general, the bottleneck of fMRI is temporal resolution, which is typically on the order of seconds, i.e., one- or two-order of magnitude larger than the duration of the cognitive processes investigated. Magneto-Encephalography [MEG: Hämäläinen *et al.*, 1993] is a technique that permits high-quality analysis of neural currents at the millisecond timescale. It allows a temporal resolution that is comparable to ERP, but a much finer-grained spatial resolution. Other neuroimaging techniques that provide unique advantages include near-infrared spectroscopic imaging (NIRSI) that measures neuronal activity in the brain from the top of the scalp, magnetic resonance spectroscopy (MRS) and chemical shift imaging (CSI) that permit the measurement of certain metabolites as an index of neuronal effort, fatigue, or decline. Finally, various combinations of imaging methods are used to maximize the benefits of each method by coregistering their activities (fMRI-MRS, fMRI-PET, fMRI-EEG, etc.).

4. "I" for "I": The Human-Like Self in a Machine

Modern computer technology, together with the state-of-the-art in artificial intelligence, makes it possible to create intelligent teachable agents compatible with humans in their core learning abilities, including the abilities to acquire language,

master creativity skills, and develop general and specialized intelligence. As a result, machines possessing general and specialized intelligent capabilities at a human level and above are anticipated to emerge in the future, becoming useful members of the human society. This challenge has an integral nature, and its solution will occur upon reaching the critical mass discussed above, as a result of a self-sustained chain reaction of cognitive growth rather than a sequence of incremental steps in development of cognitive architectures. It is therefore vital to identify the key elements of the critical mass and to find out which of them are missing in existing cognitive architectures.

In our view, the most vital key element of the critical mass that is still missing in modern cognitive architectures is the complex of functional characteristics associated with the human notion of "I", or the self. Here are some arguments supporting this point of view.

The sense of "I" or self is the central concept in techniques known in educational science as self-regulated learning (SRL), and it is also a powerful device used in human metacognition in general. Self-regulation refers to the degree to which a learner is a metacognitively, motivationally, and behaviorally active participant in his or her learning process [Zimmerman, 2002]. SRL is a critical strategic thinking process for supporting students' abilities to learn and solve problems. The concept of SRL plays a central role in modern educational science. In general, SRL involves a complex set of techniques and strategies employed by learners for deliberate regulation of their learning processes [Winne and Perry, 2000; Winne and Nesbit, 2009]. According to Zimmerman [1990; 2000; 2008], SRL includes three phases: (i) *Forethought*: understanding the task, setting goals and attitudes, selecting strategies, planning steps. (ii) *Performance*: executing the plan, trying out strategies under self-monitoring and self-control. (iii) *Reflection*: self-evaluation, causal attribution of outcomes, conflict resolution, adaptation, etc. The concept of self is critically involved here.

Furthermore, it is known that the cognitive leap in human development at the age of three to four years, during which a large complex of higher cognitive abilities become available (most of which are vital for subsequent cognitive growth), is generally associated with the emergence of the adult kind of self in a child [Bartsch and Wellman, 1995; Moore and Lemmon, 2001]. Therefore, it seems that having a human-like self is a prerequisite for an artifact to be able to learn like a human.

Yet, the nature of the human self is poorly understood and seldom modeled computationally. Most designers of popular cognitive architectures would agree that their implemented agents lack any sense of self, and they do not consider this a drawback. The reason is that the notions of a self and self-awareness are understood in the modern artificial intelligence literature in a limited sense. E.g., "The self" may refer to the robot's body, or to the running software, or to the set of variables under homeostatic control by the agent, or to the agent as a whole contrasted with other agents or the environment (some agents have this kind of self). These basic notions of the self play the grounding role with respect to the more elaborate concepts of personhood [Damasio, 1999].

The essence of the human sense of "I" is, in fact, simple and distinct from them: it can be understood as an idealized abstraction of a subject, who is the owner of experiences and the author of volitions [*minimal self*: Gallagher, 2000; *conscious self*: Samsonovich and Nadel, 2005]. These ideas were taken as the basis for the design of the self-aware cognitive architecture GMU-BICA [Samsonovich and De Jong, 2005; Samsonovich *et al.*, 2009].

The sense of "self" is also pivotal in understanding the computational nature of another hallmark of human behavior: consciousness. A favorite of philosophical debates for centuries, the nature of consciousness had been discussed in the past within AI research [e.g., McCarthy, 1999; Sloman and Chrisley, 2003], and has become a prominent research field in cognitive psychology and cognitive neurosciences [Baars, 1997; Koch 2004]. Despite its elusive nature, several theories have been advanced to quantify and model consciousness [Balduzzi and Tononi, 2008; Dehaene *et al.*, 1998]. In the field of cognitive architectures, there have been some speculations on the nature of consciousness in ACT-R [Anderson, 2007] while the LIDA cognitive architecture was explicitly designed to incorporate a computational definition of consciousness [Franklin, 2007], cf. Sloman [2010].

Yet at the same time the concept of a self remains poorly understood by major AI scholars [e.g., Sloman, 2008]. The problem appears to be of a terminological nature: it is necessary to pinpoint the precise notion of the self that is relevant to making progress in AI before a discussion of this concept can turn fruitful. To conclude this section, we point that recently good progress was made in linking the notions of self, consciousness and cognitive architectures by Robert Van Gulick, based on the development of his higher-order global state (HOGS) concept [Van Gulick, 2001; 2003]. This theory could provide a good basis for future metacognitive architectures.

5. "C" for Challenge: Bridging Models and Data Together

5.1. *Integrated repository*

The increase in computational power and software sophistication in recent decades has enabled the growth of increasingly complex models of cognitive and brain functions. Increasingly complex models *per se*, however, are not a guarantee of progress. Models can be developed at different levels of abstraction, making their relationship to each other obscure until clear links between levels are established. Modeling paradigms also tend to specialize to particular classes of tasks for which they are well suited, assuming but seldom establishing their applicability and relevance to other types of tasks. Sets of mechanisms and representations are often posited and bundled together, making credit assignment of successes to individual components difficult to perform, and generalization difficult. While models accumulate, true cumulative progress remains elusive.

To solve the problem, we propose to create an integrated public repository of BICA. The objective is to identify the necessary means to achieve greater rates of convergence and incremental progress in cognitive modeling through the use of a

shared repository of computational cognitive models, experimental tasks, and performance data. This repository would serve multiple complementary purposes.

First, an integrated repository would facilitate direct comparison of different models. The development and study of cognitive models has been ongoing for decades, yet it is difficult to see how different models map onto each other, what features or components are missing, and what progress has been made. Different modeling communities speak different languages and largely ignore each other. Detailed models and data are seldom available in a comparable format, making direct model comparisons partial at best and tendentious at worst. Therefore, we need to develop a common language for the description of modeling paradigms to achieve an understanding of how they relate to each other. An integrated repository would promote the use of common tasks and data sets as benchmarks for each subfield. It would lead to the development of shared, widely accepted comparison metrics.

As a first step toward these goals, we created a comparative table of main cognitive architectures using collective efforts of many researchers involved in their design, study, and usage. The current version of the comparative table is available at http://bicasymposium.com/cogarch. Reaching a cross-community agreement on the structure and the content of the table at this level will be a step allowing us to move down to details and forward — to further goals.

A second, practical, purpose of the repository is to provide a centralized resource that modelers, students, and teachers can access when they want to start a modeling research project. The repository should facilitate finding all the available models for one's needs and purposes. This includes source code, executable, documentation, papers, and support community. Finding all the existing models for a given task or problem can be difficult since while some tasks are well identified, others can arise in many different forms. The repository would also make available all relevant behavioral or neuroscience data for a given task. The raw data rather than the aggregate analyses would be provided in publications as additional constraints for the development of increasingly refined models. Finally, together with data the repository would make available an implemented version of the corresponding experimental tasks. Too much time (as much as half by some estimates) in modeling projects is spent (re)implementing and connecting to task environments. Often different modelers abstract away from a common task and thus prevent models from being directly comparable.

Thus, the repository would provide an immediate and organized way to access an overview of relevant information, especially key findings of specific subfields for which to develop and validate models. Making available all relevant results would promote broad and integrated rather than partial and selective accounts. Conversely, the repository would provide a consistent and comparable record of activity for the various modeling frameworks. This would provide an archival record of the range of coverage and would highlight the core focus of each framework. It would also encourage keeping models updated to keep credit for successive versions of the framework.

Another function of the repository is to enable the re-use and integration of models. That would in turn promote consistency in parameters across models, and discourage excessive (i.e., *post hoc*) parameter fitting and in favor of consensus values. Similarly, re-use and integration of models would promote ontological consistency in domain representation. The availability of standard ways of encoding knowledge for specific domains would enable the development of more complex, comprehensive models validated over broader range of findings.

A practical benefit of an integrated repository would be to encourage the development of modeling tools and standards. Developing modeling tools (e.g., for model editing or parameter search) is a rather esoteric niche with little benefits. Making them available to a broad community would benefit the community through improved productivity. The repository would also promote the development of standards, such as for the integration of models and tasks environments. This would raise productivity as well as provide additional constraints on models.

5.2. *Practical considerations*

While similar repositories have proven successful in fields such as biology and physics, a major practical issue is how to bootstrap them to the point where they become self-sustaining. A key enabling factor is to give proper academic credit for uploading materials. This would require limiting submissions to materials associated with published papers, or subjecting submissions to independent peer review. Full descriptions of models can then be referenced with a DOI system or counted as online publication as in Scholarpedia (which includes a smart revision system).

To encourage submission, making models and data available in repository should be made a condition of publication and/or funding (as it happens in other fields). This can easily be done for specialized conferences (e.g., NIPS, ICCM, BICA). Another incentive is that making behavioral data available for a given task will establish it as a de facto benchmark for its subfield. This will lead to a convergence towards a standardized set of tasks that will keep expanding rather than remain static and thus subject to be gamed, as is often the case with fixed benchmarks.

Tying the repository into an external computational system would allow users to make use of that system with no extra investment in effort. Examples of such external systems include simulation systems (e.g., Unreal Tournament and CASTLE: https://project.setcorp.com/castle/), model running, and parameter optimization system (e.g., MindModeling.org) and experiment system (e.g., Eprime). This would also enforce some code-compliance and standardization policies.

Irrespective of popularity, practical issues remain in making a repository successful. Simply uploading tasks and model code is not enough. A number of issues should be considered. Most fundamentally, a standard interface between cognitive models and task environments is needed to assure portability across tasks and models. Tasks and models could only be included in the repository when they are compatible, ensuring interoperability. If both tasks and models comply with the

interface, both scientific (principled model comparison, separation between task and model) and technical (re-usability, productivity) goals will be enhanced and the exponential growth associated with systems embracing common standards (e.g., the Internet, the Personal Computer) will then be possible. The primary scientific obstacle to such a common interface is to agree upon a common level of description across models. The alternative is to adopt a multi-level model approach that integrates models across multiple grain scales.

Another maintenance issue is how computational models need to be updated and kept current. Developers should have incentives to maintain their code up-to-date to claim cumulative credit from models developed under previous versions of their framework. The main issue is how much standardization should be required (e.g., fixed parameters, common knowledge representation) for a framework to claim an integrated account across the models that it supports.

Most practically, infrastructure funding for the repository should be provided by a funding source, e.g., Department of Defense (DOD) research agencies, DARPA, NSF, or private foundations. The alternative is incremental funding through individual projects contributing ancillary development to the repository, which would likely result in slower, piecewise development. Even if centralized, repository development should be focused on the modeler's needs through informal pools and surveys to make sure that it corresponds to actual developmental patterns and supports the modeling activity.

6. "A" for Architecture: Where it all Comes Together

6.1. *Scaling up complexity*

The ultimate aim of BICA is to reproduce intelligent human-like behavior. This has been the realm of cognitive psychology, a field within which there exists a long and successful tradition of computational modeling. A subgoal in this challenge is to create a computational replica of the human general learning ability. The robustness and scalability of learning should define the success criterion.

Even in this field, the convergence between brain sciences and AI is making it possible to create ambitiously complex cognitive models. These models are shedding new light on the origins and nature of human cognitive processes, and are valuable tools in exploring the computations performed by biological intelligent systems. The progress that the brain sciences have made in describing and understanding specific brain circuits facilitates their integration into a consistent system. For this reason, the last few years have seen the appearance of unusually large models that are both biologically grounded and capable of complex behaviors.

There are now many examples of biologically grounded systems that are functionally rich and exhibit a remarkable capacity for performing complex tasks. For instance, the Leabra architecture [O'Reilly and Munakata, 2000] is a unified framework where separate modules for vision, action, and working memory (the latter including the basal ganglia gating system) can be connected into a single system.

Those modules use similar, consistent principles but provide different functionality by using distinct parameter sets, allowing functional differentiation between neural areas while preserving architectural unification. As a result, these integrated models can realistically perform complex tasks (such as the 1-2-AX which requires two working memory loops for maintaining information [O'Reilly and Frank, 2006]) that just a few years ago would have been considered beyond the capabilities of then-current neural networks.

The possibility of interpreting micro-level neural computations in terms of macro-level symbolic algorithms also provides a natural means to scale up to more abstract and treatable forms of computations. One interesting example is offered, again, by the study of the basal ganglia. Advances in the understanding of the computational properties of the dopamine signal have led to the identification of certain parts of this circuit with the "actor" component in the actor-critic architecture [Joel *et al.*, 2002]. This has led to a closer computational exploration of the properties of the "actor", and in particular on the nature of the "actions" that the circuit can perform. These "actions" have been successfully interpreted as the opening or closing of specific "gates" that control the flow of information to working memory [Frank *et al.*, 2001]. Building from this assumption, several authors have suggested that the basal ganglia can be ultimately interpreted as an executor of conditional IF-THEN rules, like in a traditional production system [Brown *et al.*, 2004; Stocco *et al.*, 2010]. This identification has created an unexpected bridge between the neural circuitry of the brain and a well-known and widely-adopted control model that is ubiquitous in cognitive science and in AI research. At the same time, neural correlates of cognitive functions associated with production rules are probably not limited to basal ganglia and are likely to involve the medial temporal and the prefrontal cortices, among other brain structures.

In summary, advances in brain research have created novel and unexpected connections with AI, at different levels of analysis. On the one hand, increased understanding of brain physiology and microcircuitry have made it possible to constrain neural models and explore new and novel ways of neural information processing. On the other hand, unexpected similarities have been found between the behavior of certain neural populations and the terms in a number of AI learning algorithms. Finally, progress has been made on how to understand the behavior of brain circuits in terms of higher-level components and approximations, opening up the possibility for large-scale simulations.

6.2. *Convergence in architectures*

Cognitive architectures can be thought of as providing the basic computational primitives for an artificial mind, upon which specific task behaviors can be constructed [Anderson, 1983]. Research in this field dates back to Allen Newell, who outlined the first general problem-solving system [Ernst and Newell, 1969] and introduced the use of production systems as a model of cognitive control [Newell, 1973a]. Over the years, many different cognitive architectures have been proposed [Anderson, 1983; Thibadeau

et al., 1982; Laird and Newell, 1983; Meyer and Kieras, 1997; Just and carpenter, 1992; Mitchell *et al.*, 1989]. Research in this field has produced multiple, often incompatible approaches, that lately seem to be converging under biological constraints.

Research on cognitive architectures has been remarkably important in both AI and cognitive psychology. In AI, it has been one of the few research fields that have stressed the need for general intelligence, as opposed to research on problem-specific algorithms and techniques. It has also kept a healthy focus on the design of complex systems and the re-use of knowledge representations across different domains.

Traditionally, cognitive architectures in AI do not specifically aim at reproducing any biological properties of intelligence, but only at creating artificial agents that can behave and act intelligently and successfully operate in an environment. Thus, AI-oriented cognitive architectures differ in design and are typically evaluated on the grounds of what they can possibly achieve. Occasionally, however, their design choices are influenced by the observation of the architecture of natural systems. In recent years, for instance, this has suggested the integration of perception and action into a connected system [Brooks, 1986; Bell, 1999].

In the domain of cognitive neuroscience and cognitive psychology, the existence of multiple incompatible architectures has been regarded as more problematic than in AI. The reason is obvious: A truly "cognitive" architecture like ACT-R [Anderson, 2007] or EPIC [Meyer and Kieras, 1997] not only aims at producing intelligence behavior across a variety of tasks, but also claims a strict correspondence with the processes occurring in the human mind. In this field, two different architectures can both be wrong, but can never both be right. The traditional behavioral measures of cognitive psychology, e.g., reaction times and accuracy, are often not sufficient to distinguish between the two. For instance, the ACT-R and EPIC architectures divide on the existence of a central cognitive bottleneck, with ACT-R claiming its existence and predicting serial cognitive processing, and EPIC negating it and predicting parallelism. A number of behavioral experiments and computer simulations were run [Hazeltine *et al.*, 2002; Anderson *et al.*, 2005] showing that the observed pattern of reaction time data could, in fact, be explained by both architectures. Similar difficulties in testing different architectures, coupled with their complexity and large number of free parameters, have indeed made some psychologists suspicious of the approach altogether [Roberts and Pashler, 2000].

It is not surprising that cognitive architectures have welcomed the progresses in the neurosciences. Assumptions that are untestable at the behavioral level become testable at the biological level. The possibility of investigating the neural hardware gives a new meaning to the term "architecture". In fact, architectures such as ACT-R and 4CAPS [Just and Varma, 2007] have moved from explaining behavioral phenomena to modeling neuroimaging data, and Soar [Laird, 2008] has moved from simply performing complex tasks to achieving a degree of psychological fidelity. The result is a convergence between architectures at the organizational (modular) level, if not at the mechanistic level. This convergence can also happen between architectures at different levels of description, e.g., ACT-R and Leabra [Jilk *et al.*, 2008].

A third forcing factor that is pushing cognitive architectures towards BICA is the broadening of the scope of such architectures. Traditionally, cognitive architectures were limited to tasks such as reasoning, memory, learning, and language processing. More recently, interest has been growing for phenomena such as creativity, life-long learning, and consciousness [Samsonovich *et al.*, 2009; Franklin, 2007]. All of these phenomena, like consciousness, are typically attributed to and studied in biological systems. Life-long learning in humans is of particular interest, as it has no counterpart in lower animals or in artifacts and plays the pivotal role in the emergence of human intelligence.

In summary, cognitive architectures can benefit from biological constraints in at least three ways. They can use those constraints to distinguish between realistic and unrealistic assumptions and better model behavioral data; they can get insight on the design choices that underpin the intelligence and adaptivity of natural agents; and they finally can take inspiration to attack the computational nature of some hardcore and hitherto unreachable properties of biological systems.

7. Concluding Remarks

This closing part reflects upon the future of BICA. It seems that the conclusion could be somewhat similar to Newell's [1973b] review. Compared to Newell's earlier position, we can at this point not just point to an architecture as a possible answer, but also point to what will be needed in architectures. In this paper we tried to accomplish an unusual task: to briefly and quickly discuss the current state-of-the-art in BICA research, and to note what we think are the four most important forces driving the field of BICA at this point in time. We discussed the importance of critical mass, chain reactions, scalability and scaling laws, tests and metrics, and, of course, steps for a roadmap for future development, improvements and applications of BICA. In this discussion, we have outlined the deep connection and cross-fertilization between the study of general artificial intelligence and the study of how the brain enables cognition. In the past 20 years, advances in one field have found resonance in the other, and there is increasing room for further investigations towards an understanding of the general computational basis of intelligence through its biological roots. Although notable progress has been made, there are a number of possible future directions where the integration between brain and computer sciences is likely to expand. We would like to finish this manifesto by a short list of promising future directions in BICA research.

7.1. *Cognitive flexibility*

A hallmark of human intelligence is flexibility, i.e., the capacity of strategically allocating or diverting resources to a task. That capacity is the first and most important criterion in the proposed Newell test for a theory of cognition [Anderson and Lebiere, 2003]. In contrast, most AI systems are brittle and inflexible, unable to

generalize beyond their pre-programmed abilities to the variety of tasks and needs that populate real-world human experiences.

At an even deeper level, flexibility is a property of the human brain. A network of interconnected brain region is recruited every time we perform a task. However, exactly which regions are recruited and why depends on task properties, such as whether it is novel or practiced, complex or difficult, perceptual or abstract, and so on. It is not clear exactly how different regions are selected, and on the basis of which characteristics, but we do know that regions are dynamically recruited, with new areas becoming active as the old ones are outstripped of their resources [Prat *et al.*, 2007]. Very different computational explanations have been given (compare, for instance, Just and Varma, [2007], with Anderson *et al.*, [2008]). Again, insight on the necessary computations required to perform these task reconfigurations will not only shed light on some fundamental underpinnings of human intelligence, but also allow us to reproduce it in our machines.

7.2. *Life-long learning and plasticity*

AI has developed important algorithms that deal with learning in specific domains. Real-life learning in humans, however, is substantially more complex. A first notable difference is that human learning spans a number of years, with new facts being learned and continuously assimilated within existing knowledge. A second difference is that real-life learning is strategic, as humans are able to decide when and what to learn. Finally, humans surpass machines in their metacognitive abilities, being able to reflect on their own learning abilities and how to improve their own performance.

Developing more realistic long-term learning machines could have a number of practical applications in education. Cognitive architectures yielded an important result in educational technology with the development of intelligent tutoring systems, where cognitive models are created and dynamically updated to match the student's skills [Ritter *et al.*, 2007]. More realistic cognitive models can largely improve their efficacy, and some attempts have been made to extend the tutors with the use of neuroimaging data [Anderson *et al.*, 2010]. Finally, the most promising contribution from BICA to educational technologies is expected in the field of metacognitive tutoring systems [Azevedo and Witherspoon, 2009].

In the brain, learning is ultimately due to synaptic plasticity. However, plasticity encompasses other changes in the brain structure, such as development and aging. It is entirely possible, therefore, that BICA models could be used to predict a child's scholastic achievements, or to counter the effects of aging on cognitive abilities in the elderly. The most interesting task, however, is to reproduce in a BICA the process of human cognitive growth from a child to an adult.

7.3. *Emotions*

For a long time emotions had been considered entirely private experiences, laying beyond the tools of scientific investigation or, even more, AI. With hindsight, we can

see how progress on BICA was curbed by not taking into account this large and fundamental part of the human experience.

Fortunately, the scientific advances of the past 15 years have brought emotions within the spotlight of research and modeling. This was partly due to the realization that emotions are not so neatly separated from cognition as was originally thought. Not only can the circuits underlying specific emotions be localized in the brain, but it has been shown that damage to these circuits crucially affects human behavior, compromising even high-level abilities such as planning and decision-making [e.g., Damasio, 1994]. These considerations have pushed a number of researchers to investigate the computational leverage of emotions, and attempt to reproduce them in machines [Picard, 2000], see also Ventura [2010].

Despite these remarkable progresses, much remains to be done to reach a consensus in the field. Many cognitive architectures (such as ACT-R, Soar, and EPIC) still lack emotional capabilities, and it is not clear how they will be integrated in the future, while certain steps are being taken toward the development of this understanding.

In addition, understanding and displaying emotions is a crucial part of social interactions. Therefore, understanding them within a computational framework is a necessary step for developing social agents that can interact naturally with humans.

7.4. *Extending the brain and the mind*

This article has stressed the goal of understanding the nature of brain computations as a means to produce artificial, complex intelligent agents. Here and there, we have shown how the benefits can be reciprocal; for instance, reinforcement learning algorithms have provided the computational framework to understand one crucial contribution of dopamine neurons in the midbrain. Perhaps, the field where BICA can give the most immediate practical contribution to health sciences is that of brain prosthetics.

A fixture in science fiction, brain prosthetics consist in the ability to restore function in a damaged brain by connecting it to an artificial device that supplies the missing computations [Schwartz, 2004]. So far, success in this field has been limited to perceptual and motor abilities, such as Dobelle's visual and Schwartz's motor prosthetics [Vellist *et al.*, 2008]. But as our understanding of the computational nature of cognitive processes and intelligence grows, restoring higher-level functions such as language, learning capabilities, and control functions will become possible. Experiments on creating artificial prosthesis for the hippocampus, so that episodic memories can be restored in lesioned rats, have been performed [Song *et al.*, 2009]. The BICA approach is unique in its integrative approach, where the contribution of different circuits is framed within the functions of the entire system. As such, it might be the best approach to make sense of the contributions that higher-level regions make to intelligent behavior, and to restore their functions. It is therefore possible that one day we will be able to extend the substrate of our minds into a machine.

Acknowledgments

We are grateful to Drs. Frank Ritter, David Noelle, and many others for inspiring discussions of the ideas reflected in this paper. Christian Lebiere and Andrea Stocco were supported by grant FA9950-08-1-0404 from the Air Force Office of Scientific Research. Alexei Samsonovich was supported by a minigrant from the George Mason University Center for Consciousness and Transformation.

References

Albus, J. S., Bekey, G. A., Holland, J. H., Kanwisher, N. G., Krichmar, J. L., Mishkin, M., Modha, D. S., Raichle, M. E., Shepherd, G. M. and Tononi, G. [2007] "A proposal for a decade of the mind initiative," *Science* **317**, 1321.

Albus, J. S. [2010] "Reverse engineering the brain," *International Journal of Machine Consciousness* **2**(2), 193–211.

Anderson, J. R. [1983] *The Architecture of Cognition* (Lawrence Erlbaum Associates, NJ).

Anderson, J. R. [2007] *How Can the Human Mind Occur in the Physical Universe?* (Oxford University Press, New York, NY).

Anderson, J. R., Betts, S. A., Ferris, J. L. and Fincham, J. M. [2010] "Neural imaging to track mental states while using an intelligent tutoring system," *Proceedings of the National Academy of Sciences* **107**, 7018–7023.

Anderson, J. R., Fincham, J. M., Qin, Y. and Stocco, A. [2008] "A central circuit of the mind," *Trends in Cognitive Sciences* **12**, 136–143.

Anderson, J. R., Taatgen, N. A. and Byrne, M. D. [2005] "Learning to achieve perfect time-sharing: Architectural implications of Hazeltine, Teague, and Ivry," *Journal of Experimental Psychology: Human Perception and Performance* **31**, 749–761.

Anderson, J. R. and Lebiere, C. L. [2003] "The Newell test for a theory of cognition," *Behavioral & Brain Sciences* **26**, 587–637.

Azevedo, R. and Witherspoon, A. M. [2009] "Detecting tracking, and modeling self-regulatory processes during complex learning with hypermedia," *Biologically Inspired Cognitive Architectures: Papers from the AAII Fall Symp.* [2009], pp. 16–26.

Baars, B. J. [1997] *In the Theater of Consciousness* (Oxford University Press, New York).

Balduzzi, D. and Tononi, G. [2008] "Integrated information in discrete dynamical systems: Motivation and theoretical framework," *PLoS Computational Biology* **4**(6).

Babb, S. J. and Crystal, J. D. [2006] "Episodic-like memory in the rat," *Current Biology*, **16**, 1317–1321.

Bartsch, K. and Wellman, H. M. [1995] *Children Talk About Mind* (Oxford University Press, New York).

Barto, A. G. [1995] "Adaptive Critics and the Basal Ganglia," in *Models of Information Processing in the Basal Ganglia*, J. C. Houk, J. L. Davis and D. G. Beiser (eds.) (MIT Press, Cambridge), pp. 215–232.

Bell, A. J. [1999] "Levels and loops: The future of artificial intelligence and neuroscience," *Philosophical Transaction of the Royal Society London*, **B354**(1392), 2013–2020.

Brooks, R. [1986] "A robust layered control system for a mobile robot," *IEEE Journal of Robotics and Automation* **2**, 14–23.

Brown, J. W., Bullock, D. and Grossberg, S. [2004] "How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades," *Neural Networks* **17**, 471–510.

Dally, J. M., Emery, N. J. and Clayton, N. S. [2010] "Avian theory of mind and counter espionage by food-caching western scrub-jays (aphelocoma californica)," *European Journal of Developmental Psychology* **7**(1), 17–37.

Damasio, A. R. [1994] *Descartes' Error* (Putnam, New York).
Damasio, A. R. [1999] *The Feeling of What Happens* (Harcourt, New York).
Dehaene, S., Kerszberg, M. and Changeux, J.-P. [1998] "A neuronal model of a global workspace in effortful cognitive tasks," *Proceedings of the National Academy of Sciences* **95**, 14529–14534.
Ernst, G. and Newell, A. N. [1969] *GPS: A Case Study in Generality and Problem Solving* (Academic Press, New York).
Frank, M. J., Loughry, B. and O'Reilly, R. C. [2001] "Interactions between frontal cortex and basal ganglia in working memory: A computational model," *Cognitive, Affective & Behavioral Neuroscience* **1**, 137–160.
Franklin, S. [2007] "A foundational architecture for artificial general intelligence," *Frontiers in Artificial Intelligence and Applications* **157**, 36–54.
Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E. and Penny, W. D. [2006] *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (Elsevier, London).
Fyhn, M., Hafting, T., Treves, A., Moser, M. B. and Moser, E. I. [2007] "Hippocampal remapping and grid realignment in entorhinal cortex," *Nature* **446**(7132), 190–194.
Gallagher, S. [2000] "Philosophical conceptions of the self: Implications for cognitive science," *Trends in Cognitive Science* **4**, 14–21.
Hafting, T., Fyhn, M., Molden, S., Moser, M. B. and Moser, E. I. [2005] "Microstructure of a spatial map in the entorhinal cortex," *Nature* **436**(7052), 801–806.
Hallett, M. [2000] "Transcranial magnetic stimulation and the human brain," *Nature* **406**, 147–150.
Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J. and Lounasmaa, O. V. [1993] "Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain," *Review of Modern Physics* **65**, 413–497.
Hazeltine, E., Teague, D. and Ivry, R. B. [2002] "Simultaneous dual-task performance reveals parallel response selection after practice," *Journal of Experimental Psychology: Human Perception and Performance* **28**, 527–545.
Hopfield, J. J. [1982] "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences* **79**, 2554–2558.
Hunt, G. R. [1996] "Manufacture and use of hook-tools by new caledonian crows," *Nature* **379**, 249–251.
Jilk, D. J., Lebiere, C., O'Reilly, R. C. and Anderson, J. R. [2008] "SAL: An explicitly pluralistic cognitive architecture," *Journal of Experimental & Theoretical Artificial Intelligence* **20**(3), 197–218.
Joel, D., Niv, Y. and Ruppin, E. [2002] "Actor-critic models of the basal ganglia: New anatomical and computational perspectives," *Neural Networks* **15**, 535–547.
Just, M. A. and Carpenter, P. A. [1992] "A capacity theory of comprehension: Individual differences in working memory," *Psychological Review* **99**, 122–149.
Just, M. A. and Varma, S. [2007] "The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition," *Cognitive, Affective and Behavioral Neuroscience* **7**, 153–191.
Koch, C. [2004] *The Quest for Consciousness* (Roberts & Company, Englewood, CO).
Laird, J. E. [2008] Extending the soar cognitive architecture, Artificial General Intelligence Conference, Memphis, TN.
Laird, J. and Newell, A. [1983] *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (Morgan Kaufmann Publishers Inc., San Francisco, CA), pp. 771–773.
Langley, P., McKusick, K. B., Allen, J. A., Iba, W. F. and Thompson, K. [1991] "A design for the ICARUS architecture," *ACM SIGART Bulletin* **2**, 104–109.

Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. A., Pappata, S., Molko, N. and Chabriat, H. [2001] "Diffusion tensor imaging: Concepts and applications," *Journal of Magnetic Resonance Imaging* **13**, 534–546.

Licklider, J. C. [1968] "The computer as a communication device," *Science and Technology*, April.

McCarthy, J. [1999] "Making robots conscious of their mental states," *Machine Intelligence* **15**, 3–17.

McNaughton, B. L., Barnes, C. A., Gerrard, J. L., Gothard, K., Jung, M. W., Knierim, J. J., Kudrimoti, H., Qin, Y., Skaggs, W. E., Suster, M. and Weaver, K. L. [1996] "Deciphering the hippocampal polyglot: The hippocampus as a path integration system," *Journal of Experimental Biology* **199**, 173–185.

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I. and Moser, M. B. [2006] "Path integration and the neural basis of the 'cognitive map'," *Nature Reviews Neuroscience* **7**(8), 663–678.

Meyer, D. E. and Kieras, D. E. [1997] "A computational theory of executive cognitive processes and multiple-task performance. 1. Basic mechanisms," *Psychological Review* **104**, 3–65.

Mills, D. L., Plunkett, K., Prat, C. S. and Schafer, G. [2005] "Watching the infant brain learn words: Effects of vocabulary size and experience," *Cognitive Development* **20**, 19–31.

Minsky, M. L. and Papert, S. A. [1969] *Perceptrons* (MIT Press, Cambridge, UK).

Mitchell, T. M., Allen, J., Chalasani, P., Cheng, J., Etzioni, O., Ringuette, M. and Schlimmer, J. [1989] "Theo: A framework for self-improving systems," in *Architectures for Intelligence*, K. VanLehn (ed.) (Erlbaum, 1990).

Moore, C. and Lemmon, K. (eds.) [2001] *The Self in Time: Developmental Perspectives* (Lawrence Erlbaum Associates, Mahwah, NJ).

Newell, A. N. [1973a] "Production systems: Models of control structure," in *Visual Information Processing*, W. G. Chase (ed.) (Academic Press, New York), pp. 526–547.

Newell, A. N. [1973b] "You can't play 20 questions with nature and win," in *Visual Information Processing* W. G. Chase (ed.) (Academic Press, New York), pp. 283–308.

Niv, Y., Duff, M. O. and Dayan, P. [2005] "Dopamine, uncertainty and TD learning," *Behavioral and Brain Functions* **1**, 6.

O'Keefe, J. and Dostrovsky, J. [1971] "The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely moving rat," *Brain Research* **34**, 171–175.

O'Reilly, R. C. and Munakata, Y. [2000] *Computational Explorations in Cognitive Neuroscience* (MIT Press, Cambridge, MA).

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.

Pepperberg, I. M. and Gordon, J. D. [2005] "Number comprehension by a grey parrot (psittacus erithacus), including a zero-like concept," *Journal of Comparative Psychology* **119**, 197–209.

Picard, R. W. [2000] *Affective Computing* (MIT Press, Cambridge, MA).

Plaut, D. C. [2002] "Graded modality-specific specialization in semantics: A computational account of optic aphasia," *Cognitive Neuropsychology* **19**, 603–639.

Posner, M. I. and Raichle, M. E. [1994] *Images of Mind* (Scientific American Books, USA).

Prat, C. S., Keller, T. A. and Just, M. A. [2007] "Individual differences in sentence comprehension: A functional magnetic resonance imaging investigation of syntactic and lexical processing demands," *Journal of Cognitive Neuroscience* **19**, 1950–1963.

Ritter, S., Anderson, J. R., Koedinger, K. R. and Corbett, A. [2007] "Cognitive tutor: Applied research in mathematics education," *Psychonomic Bulletin & Review* **14**, 249–255.

Roberts, S. and Pashler, H. [2000] "How persuasive is a good fit? A comment on theory testing," *Psychological Review* **107**, 358–367.

Rosenblatt, F. [1958] "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review* **65**, 386–408.

Rumelhart, D. E. and McClelland, J. L. [1986] *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, MA).

Samsonovich, A. V. and De Jong, K. A. [2005] "Designing a self-aware neuromorphic hybrid," *AAAI-05 Workshop on Modular Construction of Human-Like Intelligence.*

Samsonovich, A. and McNaughton, B. L. [1996] "Path integration and cognitive mapping in a continuous attractor neural network model," *The Journal of Neuroscience* **17**, 5900–5920.

Samsonovich, A. V. and Nadel, L. [2005] "Fundamental principles and mechanisms of the conscious self," *Cortex* **41**(5), 669–689.

Samsonovich, A. V., De Jong, K. A. and Kitsantas, A. [2009] "The mental state formalism of GMU-BICA," *International Journal of Machine Consciousness* **1**, 111–130.

Savelli, F. and Knierim, J. J. [2010] "Hebbian analysis of the transformation of medial entorhinal grid-cell inputs to hippocampal place fields," *Journal of Neurophysiology* **103**, 3167–3183.

Schultz, W. [1998] "Predictive reward signal of dopamine neurons," *Journal of Neurophysiology* **80**, 1–27.

Schultz, W., Apicella, P. and Ljungberg, T. [1993] "Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task," *Journal of Neuroscience* **13**, 900–913.

Schwartz, A. B. [2004] "Cortical neural prosthetics," *Annual Review of Neuroscience* **27**, 487–507.

Scoville, W. B. and Milner, B. [1957] "Loss of recent memory after bilateral hippocampal lesions," *Journal of Neurology, Neurosurgery and Psychiatry* **20**, 11–21.

Semendeferi, K. and Damasio, H. [2000] "The brain and its main anatomical subdivisions in living hominoids using magnetic resonance imaging," *Journal of Human Evolution* **38**, 317–332.

Sloman, A. [2008] "'The self': A bogus concept," manuscript published online at `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/the-self.html`.

Sloman, A. [2010] "An alternative to working on machine consciousness," *International Journal of Machine Consciousness* **2**(1), 1–18.

Sloman, A. and Chrisley, R. [2003] "Virtual machines and consciousness," *Journal of Consciousness Studies* **10**, 133–172.

Song, D., Chan, R. H., Marmarelis, V. Z., Hampson R. E., Deadwyler, S. A. and Berger T. W. [2009] "Nonlinear modeling of neural population dynamics for hippocampal prostheses," *Neural Networks.*

Sporns, O., Tononi, G. and Kötter, R. [2005] "The human connectome: A structural description of the human brain," *PLoS Computational Biology* **1**(4).

Stocco, A., Lebiere, C. and Anderson, J. R. [2010] "Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination," *Psychological Review* **117**, 540–574.

Sutton, R. S. [1988] "Learning to predict by the methods of temporal differences," *Machine Learning* **3**, 9–44.

Sutton, R. S. and Barto, A. G. [1998] *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Thibadeau, R., Just, M. A. and Carpenter, P. A. [1982] "A model of time course and content of reading," *Cognitive Science* **6**, 157–203.

Touretzky, D. S. and Muller, R. U. [2006] "Place field dissociation and multiple maps in hippocampus," *Neurocomputing* **69**, 1260–1263.

Van Gulick, R. [2001] "Inward and upward: Reflection, introspection and self-awareness," *Philosophical Topics* **28**(2), 275–305.

Van Gulick, R. [2003] "Maps, gaps and traps," in *Consciousness: New Philosophical Perspectives*, Q. Smith and A. Jokic (eds.) (Oxford University Press, Oxford).

Velliste, M., Perel, S., Spalding, M. C., Whitford, A. S. and Schwartz, A. B. [2008] "Cortical control of a prosthetic arm for self-feeding," *Nature* **453**, 1098–1101.

Ventura, R. [2010] "Emotions and empathy: A bridge between nature and society," *International Journal of Machine Consciousness* **2**(2), 343–361.

Winne, P. H. and Perry, N. E. [2000] "Measuring self-regulated learning," in *Handbook of Self-Regulation*, P. Pintrich, M. Boekaerts and M. Seidner (eds.) (Academic Press, Orlando, FL), pp. 531–566.

Winne, P. H. and Nesbit, J. C. [2009] "Supporting self-regulated learning with cognitive tools," in *Handbook of Metacognition in Education*, D. J. Hacker, J. Dunlosky and A. C. Graesser (eds.) (Erlbaum, Mahwah, NJ).

Witter, M. P. and Moser, E. I. [2006] "Spatial representation and the architecture of the entorhinal cortex," *Trends in Neurosciences* **29**(12), 671–678.

Yeung, N., Botvinick, M. M. and Cohena, J. D. [2006] "The neural basis of error detection: Conflict monitoring and the error-related negativity," *Psychological Review* **111**, 931–959.

Zimmerman, B. J. [1990] "Self-regulated learning and academic achievement: An overview," *Educational Psychologist* **25**, 3–17.

Zimmerman, B. J. [2000] "Attaining self-regulation: A social cognitive perspective," in *Handbook of Self-Regulation*, M. Boekaerts, P. R. Pintrich and M. Zeidner (eds.) (Academic Press, San Diego, CA), pp. 13–39.

Zimmerman, B. J. [2008] "Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects," *American Educational Research Journal* **45**(1), 166–183.