Journal of Artificial General Intelligence 4(2) 23-58, 2013
DOI: 10.2478/jagi-2013-0004

Submitted 2013-09-30
Accepted 2013-12-02

Editorial: Conceptual Commitments of AGI Systems

Haris Dindo HARIS.DINDO@UNIPA.IT
Computer Science Engineering (DICGIM)
University of Palermo
Viale delle Scienze - Edificio VI
90128 Palermo, Italy

James Marshall JMARSHALL@SARAHLAWRENCE.EDU
Computer Science Department
Sarah Lawrence College
One Mead Way
Bronxville, NY 10708, USA

Giovanni Pezzulo GIOVANNI.PEZZULO@ISTC.CNR.IT
Istituto di Scienze e Tecnologie della Cognizione (ISTC-CNR)
Via S. Martino della Battaglia, 44
00185 Roma, Italy

What are the most important design principles that we should follow to build an Artificial General Intelligence? What should be the key constituents of systems-level models of cognition and behavior?

In the target article "Conceptual Commitments of the LIDA Model of Cognition", Stan Franklin, Steve Strain, Ryan McCall, and Bernard Baars tackle these difficult problems. They propose twelve "conceptual commitments" or tentative hypotheses that form the core of the Learning Intelligent Distribution Agent (LIDA) model that they have been developing over the last ten years or so. Although the article is focused on the LIDA model, these "conceptual commitments" have much broader scope and are offered to the AGI community as specific constraints that should inform the research agenda for the realization of an Artificial General Intelligence (AGI).

The twelve specific "conceptual commitments" are of various kinds and have different degrees of importance for LIDA and AGI more generally. Some (Systems-level Modeling, Global Workspace Theory, Learning via Consciousness, Feelings as Motivators and Modulators of Learning, Transient Episodic Memory) are considered to be key for LIDA and also more broadly for AGI. These are general mechanisms of learning, memory and inference that should form the core of realistic, real-world architectures of brain and behavior. Of particular note, the authors highlight the importance (among the other things) of feeling and consciousness, which are regarded as fundamental architectural solutions to the problems of AGI. These themes, which were given minor importance in traditional cognitive (neuro)science and AI, have increasingly gained prominence in the last few years. Putting these themes at center of AGI research is a distinguishing aspect of the proposal of Franklin and collaborators.

Some other “conceptual commitments” (Biologically Inspired, Embodied or Situated Cognition, Cognitive Cycles as Cognitive Atoms, Comprehensive Decay of Representations and Memory, Asynchrony, Non-linear Dynamics Bridge to Neuroscience, Theta Gamma Coupling from the Cognitive Cycle) are very important for LIDA but not necessarily so for AGI. Most of these commitments relate to the link between AGI and neuroscience. Clearly, the question of whether or not (or to what extent) an AGI should be biologically realistic is far from settled; the authors add interesting considerations to this debate by showing the importance of biological constraints on the LIDA model.

Finally, two “conceptual commitments” (Profligacy in Learning, Consolidation) are less central to LIDA and the enterprise of AGI. Clearly, any systems-level proposal must have ancillary mechanisms that permit its functioning; deciding whether or not to elevate them to indispensable principles is again an important architectural choice, exemplified in this debate.

The target article “Conceptual Commitments of the LIDA Model of Cognition” intends to stimulate a debate in the AGI community on, first, the specific working hypotheses and design principles proposed by Franklin and co-authors; and second, and more generally, on the importance of identifying and making explicit the design principles and working hypotheses of one's own computational architecture—and even more so in the design of large scale architectures such as those targeted by the AGI community.

And indeed the target article has already generated some initial debate: the six commentaries included here have raised interesting challenges to several aspects of the proposal of Franklin and co-authors.

In their commentary, Benjamin Angerer and Stefan Schneider sound a cautionary note that AGI researchers would do well to keep in mind in developing their theories and models, namely, that the nuts-and-bolts implementation of a theory is just as important as the theoretical concepts themselves, and one must be careful to distinguish between the two. They also agree with Franklin et al. on the need to identify good general benchmarks for evaluating and comparing AGI systems, but emphasize that human cognition may have a particularly important role to play in guiding the development of these benchmarks. Finally, they point out that the integrative approach to AGI rests on a variety of implicit concepts and assumptions from cognitive psychology—assumptions that ultimately may or may not turn out to be warranted. To some extent such assumptions are necessary if we wish to build concrete models, but they nevertheless entail a certain risk. It may be necessary to rethink or abandon them to make further progress—another point of caution to keep in mind.

In his commentary, Antonio Chella applauds LIDA's commitment to consciousness as a core principle of intelligence, in contrast to other integrative cognitive architectures in which other aspects of intelligence such as problem solving, resource maximization, or integration of capabilities are regarded as the key principles. He suggests that this focus on consciousness is itself an important conceptual commitment for AGI that should be included in the list of commitments proposed by Franklin et al.

In his commentary, John Laird highlights some of the differences between the SOAR cognitive architecture and LIDA, particularly the relative emphasis that each architecture places on functional commitments versus biological or psychological commitments. He outlines a set of general functional constraints and requirements for AGI systems, and emphasizes the importance of real-time system performance. In his view, LIDA's attempt to incorporate functional, biological, and psychological constraints within a single system may be overly ambitious, at least if efficient real-time performance of the system is also a requirement of the model.

Olivier Georgeon and David Aha focus on Franklin et al.'s conceptual commitment “Cognitive Cycles as Cognitive Atoms”. This commitment is central to the LIDA model, but Franklin et al. are undecided as to its level of importance for AGI systems in general. Georgeon and Aha, however, view it as critically important, and propose an even stronger conceptual commitment called “Radical Interactionism” (RI), which recasts the basic principle of an indivisible cognitive cycle (a “cognitive atom”) in terms of sensorimotor interactions. In their view, the traditional distinction between perception and action as separate entities is unnecessary and misleading. Instead, their RI commitment subsumes perception and action into the more fundamental notion of sensorimotor interaction, which they consider to be the appropriate primitive on which to base cognitive agent architectures. They also show how the LIDA architecture could be modified to reflect this new conceptual interpretation.

In his commentary, Pei Wang reflects thoughtfully on the unique challenges faced by the field of AGI in designing general-purpose AI systems, due in part to its lack of established, agreed-upon theories and frameworks to guide research. He points out that the conceptual commitments that underlie different AGI projects may differ according to which aspects of human intelligence the projects focus on. That is, the particular research objectives of a project may determine which conceptual commitments are relevant, and these commitments may overlap only partially with those proposed by Franklin et al. He also considers the different types of challenges and risks that arise in taking an integrated versus a unified approach to AGI, and makes the important point that being able to describe some aspect of intelligence in psychological terms is not by itself sufficient justification for its implementation in an AGI system as a distinct module or mechanism.

In their commentary, Travis Wiltshire and his colleagues take issue with Franklin et al.'s apparent commitment to “disembodied embodiment”. They raise important questions regarding the LIDA architecture's level of commitment to feedback-rich agent-environment interaction in general, as well as to socially interactive capabilities in particular. More broadly, they suggest that a stronger commitment to human-like embodiment (as opposed to Franklin et al.’s less specific notion) may be necessary in order to achieve AGI's ultimate goals. Furthermore, they point out that in evaluating an AGI system, it is crucial to take into account the extent to which the system is perceived and treated by humans as conveying agency through social interactivity.

In summary, this issue of the JAGI journal hosts a stimulating debate on which design principles and “conceptual commitments” should form the foundations of large-scale systems for Artificial General Intelligence. Stan Franklin, Steve Strain, Ryan McCall, and Bernard Baars propose a rich set of important working principles that stem from their long experience with their Learning Intelligent Distribution Agent (LIDA) model. The specific principles are a matter of debate: they can be discussed, elaborated on, and called into question—and indeed the lively discussion provided by the commentators testifies that this debate has already begun. Still, the authors are, in our opinion, correct in highlighting that the time has come to distill key principles from current research in cognitive science, neuroscience, AI, machine learning, and beyond, that these principles need to be operationalized and made explicit, and that their discussion is a key research objective of the AGI (and JAGI) community. We sincerely hope that this JAGI special issue will contribute significantly to this important objective, and will stimulate the type of long-lasting debate that is crucial for the overall progress of the discipline.

Commentaries on the Target Article

General Problems of Unified Theories of Cognition, and Another Conceptual Commitment of LIDA

Benjamin Angerer BANGERER@UOS.DE
MEi:CogSci Programme
University of Vienna, Austria

Stefan Schneider STEFSCHN@UOS.DE
Institute of Cognitive Science
Albrechtstr. 28, 49076 Osnabrück
University of Osnabrück, Germany

What distinguishes the AGI approach from the initial, supposedly equally idealistic and holistic, AI approach? Why do we think that we could make any progress in our recent times? The answer to these questions is not clear, and the common tenor in AGI papers and talks is more an appeal to our joint good will. Formulated ideas on what grounds to proceed are there, yet remain rather sketchy (Adams et al., 2012).

In their target paper, Franklin et al. wonder about how the AGI community could productively interact, given that each group of researchers have their own cognitive architectures and that these architectures and their underlying Unified Theories are not easily evaluated. Their proposal is to start by discussing what they call *conceptual commitments*, the tentative hypotheses each of us bases his or her AGI research on. While this effort certainly is to be applauded, we want to briefly address some problematic aspects and relate these to the proposed, integrative paradigm.

1. Wishful Mnemonics and Implementational Caution

Some pitfalls of AI research practice have been criticized very early by Weizenbaum (1967) and McDermott (1976). For example, conceptual conflicts were predetermined when simple, computational procedures were given "cognitive names" (McDermott calls them 'wishful mnemonics'). These names, e.g. UNDERSTAND or GOAL are appealing as they feed on all our intuitions about the respective notions. As a consequence however they can obscure the real complexity of the denoted cognitive faculties and thereby impede conceptual progress.

What justifies a box in our theory overview diagram, i.e. a certain data structure or algorithm to be called e.g. 'long-term memory'? Just like with individual symbols, where Franklin et al. agree with Barsalou that they must not be amodal labels (p. 8), it is not the assigned name, but the realization of a certain functional role that justifies our box to be called 'long-term memory'. While this functional role must be specified in our theory, it is equally important to take care that it is actually realized in the implementation of the theory. Thus, it is not only necessary as a matter of principle and ontological diligence to distinguish between our theories and their implementation – as Franklin et al. do when talking about the LIDA model vs. the LIDA computational framework – but (also given the circumstances of academic software development) there is a very practical need to do so.

An often alleged asset of a comprehensive implementation is that it requires to scrutinize every detail of a theory. But the pressure to implement can also lead to quick fixes, cognitively implausible, but henceforth incorporated into the framework. In the worst case, especially if the aforementioned distinction between the theory and its implementation is not stringently upheld, such a stopgap solution could even retroact on the respective theoretical notions and intuitions.

2. The Benchmark Problem

Even with sound conceptual commitments at hand, the problem remains that we have to put our theories to the test, and that tests have to be selected that cover the spectrum of abilities we expect from an AGI agent. At the moment more or less every research group has done experimental work and can lend some empirical credence to its theories, yet only a minority of these experiments are comparable to each other (for an analysis of this problem and some suggestions cf. Adams et al., 2012).

How to check whether an architecture that claims generality (as did already the GPS) actually does the job that our intuitions point at? Seen as an integrative project, AGI would not only test individual modules, but put particular emphasis on the functioning of the whole apparatus. While the most important aim of finding a good benchmark for cognitive architectures is to achieve comparability, there is also a latent hope that by finding just the right benchmark problem we might be able to delineate general intelligence without having to explicitly specify its constituents (cf. e.g. Bringsjord and Licato, 2012). While we do neither want to exclude the possibility that general intelligence might be achieved employing vastly different means than humans nor deny the genuine interestingness of such an approach, if we want AGI to be a method of understanding the human mind (Bach, 2008) we should pay particular attention to how humans achieve the feats we are interested in.

In doing so we might even think about whether AGI should further its own methods of experimentation, not only with machines, but with humans. Considering how much progress has been made due to the early, broad investigations of the mind, such as Newell and Simon's *Human Problem Solving* (1972) and how strongly they influence AI to this day, AGI might profit from comprehensive, experimental investigations of human cognition in addition to continuing the integration of evidence from the sub-disciplines of cognitive science.

3. Integration Accepts and Confirms Compartmentalization

It might well be that the initial AI project could be so idealistic exactly because it did not foresee the huge problems general AI would have to face. Nowadays, we comfort ourselves to acknowledge this problem. The question is, how do we cope with it? One answer of AGI is to implement architectures that cover it all, based on concepts derived from psychology, neuroscience and artificial intelligence. Franklin et al. favor such an approach and suggest that the work of an AGI researcher is that of integrating and updating detail from various, scattered disciplines within "a global map". By following such an approach, LIDA implicitly subscribes to a further conceptual commitment by relying on the traditional way cognitive psychology compartmentalized cognition: perception, attention, action, thought, kinds of memory, emotion, motivation, and so forth.

An integrative approach like in the LIDA model is therefore prone to accept and reinforce such a compartmentalization. But what if we are not yet at the point where we can devise such a model of the mind, which is “correct in the broad-based sense” (p. 3)? What consequences would follow if an adjustment of one part would propagate through the whole architecture? Updating detail, like suggested by Franklin et al., could be insufficient, but substantial rethinking or even abandoning the model could become necessary. Results from the individual disciplines might even be flawed when put into the integrative context.

A practical problem for an implementation like in LIDA is that every module actually has to work to some degree. In a computational model, known conceptual “blind spots” cannot be simply left out. The intent to implement does not leave much space for essential restructuring here. However, it might well be that solutions to the hard problems of AGI (e.g. symbol grounding, implicit knowledge, motivation, creative thought) will be found here and that real progress is only possible if we put ourselves in a position that systematically allows us to take a step back, and rethink. Bringing to light and debating conceptual commitments appears to be a good starting point.

References

Adams, S.; Arel, I.; Bach, J.; Coop, R.; Furlan, R.; Goertzel, B.; Storrs Hall, J.; Samsonovich, A.; Scheutz, M.; Schlesinger, M.; Shapiro, S.; Sowa, J. 2012. Mapping the Landscape of Human-Level Artificial General Intelligence. *AI Magazine*. 33(1): 25-42.

Bach, J. 2008. Seven Principles of Synthetic Intelligence. In *Proceedings of the 2008 Conference on Artificial General Intelligence*, 63-74. IOS Press Amsterdam, The Netherlands.

Bringsjord, S., and Licato, J. 2012. Psychometric Artificial General Intelligence: The Piaget-MacGuyver Room. In *Theoretical Foundations of Artificial General Intelligence,* 25-48. Ed. Wang, P., and Goertzel, B. Atlantis Press Paris, France.

McDermott, D. 1976. Artificial Intelligence Meets Natural Stupidity. *SIGART Bulletin*. 57: 4-9.

Newell, A., and Simon, H. 1972. *Human Problem Solving*. Prentice-Hall Englewood Cliffs, NJ, USA.

Weizenbaum, J. 1967. Contextual Understanding by Computers. *Communications of the ACM.* 10(8): 474-480.

LIDA, Committed to Consciousness

Antonio Chella ANTONIO.CHELLA@UNIPA.IT
Dept. of Chemical, Management, Computer, Mechanical Engineering
University of Palermo
Viale delle Scienze, building 6
90128, Palermo, Italy

In the target paper, Franklin *et al.* propose LIDA as a reference architecture for the AGI community. In particular, the authors summarize the main commitments at the basis of the architecture.

Now, it should be noticed that the cognitive architectures commonly adopted by the AGI community are based on heterogeneous commitments. SOAR (Newell, 1990; Laird, 2012) and ACT-R (Anderson, 1983, 2007), are essentially based on the hypothesis that intelligence is strictly related to *problem solving* capabilities. In fact, these architectures are also employed in the AI community as problem solving tools.

Another competing model recently adopted by the AGI community is the AIXI agent (Hutter, 2005), based on the hypothesis that intelligence consists in the *maximization* of some utility function. AIXI implements in some sense the main principle of intelligence proposed by Wang (2006) according to which intelligence is related to optimization of resources.

Other architectures employed in the AGI community are based on the more general idea that an intelligent agent should be based on an efficient *integration* of several different cognitive capabilities, as vision, motion, problem solving, reasoning, and so on. A prototype of this large class of architectures is OpenCog, developed over the years by Ben Goertzel and his collaborators, see, e.g., (Goertzel, 2009) for an introduction.

The previously mentioned cognitive architectures essentially do not take into account, or take into account only to some limited extent, the commitment to *consciousness* at the basis of intelligence.

In fact, the relationship between consciousness and intelligence is not trivial. On the one hand, the need of consciousness for an intelligent agent is obviously taken for granted. On the other hand, it is commonly accepted that many cognitive processes that are necessary for intelligence may be unconscious and they may happen in the absence of consciousness. However, it is undeniable that consciousness is related to the broader unpredictable and less automatic forms of intelligence, as, for example, *creativity* (Boden, 2004).

Many computational models of consciousness have been recently proposed, such as the ones discussed by Shanahan (2010), Holland, Knight, and Newcombe (2007), Aleksander and Morton (2012), Arrabales (2012), Haikonen (2012), to cite just a few of them; see (Reggia, 2013) for a recent review. However, these systems, though effective and based on ideas inspired by empirical studies, are mainly focused on the investigation of particular aspects of consciousness and therefore they are not integrated in a complete cognitive architecture, with the partial exception of the model proposed by Haikonen which is still in its infancy.

Instead, the LIDA architecture is going to be a mature cognitive architecture with the peculiarity of being born and developed around the main hypothesis that intelligence is committed to consciousness. The important characteristic of LIDA is that a main mechanism for consciousness, i.e., the Global Workspace proposed by Baars (1988), is the *core* of the architecture, complemented by several different mechanisms for perception, action, reasoning, long and short term memory, and so on. And in fact, Franklin, discussing previous implementations (Franklin and Graesser, 1999; Franklin, 2003; Baars and Franklin, 2009), strongly put into evidence his commitment to consciousness.

Therefore, I admit I am quite surprised to have not found a strong initial commitment to consciousness among the commitments related to AGI and listed in the target paper. The LIDA architecture is instead a unique opportunity for the AGI community to investigate the role of consciousness as a foundational principle of intelligence.

References

Aleksander, I., and Morton, H. 2012. *Aristotle's Laptop.* Singapore: World Scientific.

Anderson, J. 1983. *The Architecture of Cognition.* Cambridge, MA: Harvard University Press.

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

Arrabales, R. 2012. Inner Speech Generation in a Video Game Non-Player Character: From Explanation to Self? *International Journal of Machine Consciousness* 4(2):367–381.

Baars, B., and Franklin, S. 2009. Consciousness is Computational: The LIDA Model of Global Workspace Theory. *International Journal of Machine Consciousness* 1(1):23 – 32.

Baars, B. 1988. *A Cognitive Theory of Consciousness.* Cambridge, MA: Cambridge University Press.

Boden, M. 2004. *The Creative Mind: Myths and Mechanisms - Second edition.* London: Routledge.

Franklin, S., and Graesser, A. 1999. A Software Agent Model of Consciousness. *Consciousness and Cognition* 8:285–301.

Franklin, S. 2003. IDA - A Conscious Artifact? *Journal of Consciousness Studies* 10(4 - 5):47 – 66.

Goertzel, B. 2009. OpenCog Prime: A Cognitive Synergy Based Architecture for Embodied Artificial General Intelligence. In *Proceedings of ICCI-09*.

Haikonen, P. 2012. *Consciousness and Robot Sentience.* Singapore: World Scientific.

Holland, O.; Knight, R.; and Newcombe, R. 2007. A Robot-Based Approach to Machine Consciousness. In Chella, A., and Manzotti, R., eds., *Artificial Consciousness.* Imprint Academic. 156 – 173.

Hutter, M. 2005. *Universal Artificial Intelligence - Sequential Decisions Based on Algorithmic Probability.* Heidelberg: Springer.

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

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

Reggia, J. 2013. The Rise of Machine Consciousness: Studying Consciousness with Computational Models. *Neural Networks* 44:112–131.

Shanahan, M. 2010. *Embodiment and the Inner Life.* Oxford, UK: Oxford University Press.

Wang, P. 2006. *Rigid Flexibility - The Logic of Intelligence.* Dordrecht: Springer.

The Radical Interactionism Conceptual Commitment

Olivier L. Georgeon OLIVIER.GEORGEON@LIRIS.CNRS.FR
Université Lyon 1
LIRIS, CNRS, UMR5205
Villeurbanne F-69622, FRANCE

David W. Aha DAVID.AHA@NRL.NAVY.MIL
Navy Center for Applied Research in AI
Naval Research Laboratory, Code 5514
4555 Overlook Ave., SW
Washington, DC 20375, USA

Abstract

We introduce *Radical Interactionism* (RI), which extends Franklin et al.'s (2013) *Cognitive Cycles as Cognitive Atoms* (CCCA) proposal in their discussion on conceptual commitments in cognitive models. Similar to the CCCA commitment, the RI commitment acknowledges the indivisibility of the perception-action cycle. However, it also reifies the perception-action cycle as *sensorimotor interaction* and uses it to replace the traditional notions of *observation* and *action*. This complies with constructivist epistemology, which suggests that knowledge of reality is constructed from regularities observed in sensorimotor experience. We use the LIDA cognitive architecture as an example to examine the implications of RI on cognitive models. We argue that RI permits self-programming and constitutive autonomy, which have been acknowledged as desirable cognitive capabilities in artificial agents.

1. Introduction

In their paper "Conceptual Commitments of the LIDA Model of Cognition", Stan Franklin, Steve Strain, and Ryan McCall invite Artificial General Intelligence (AGI) researchers to discuss which conceptual commitments are essential for AGI agents. In response to this invitation, we wish to further discuss their fourth commitment: *Cognitive Cycles as Cognitive Atoms* (CCCA). On page 9, they write: "we hesitate to propose it as important for the AGI research in general, since to our knowledge, no other system-level cognitive architecture makes such a commitment". We make this commitment in our Enactive Cognitive Architecture (ECA, Georgeon, Marshall, and Manzotti, 2013). Moreover, we extend this commitment into a more radical one called Radical Interactionism (RI).

In Section 4.4, Franklin et al. define a cognitive cycle as a cycle that begins by sampling (sense) the environment and ends by selecting an appropriate response (action), traversing various phases including perception, understanding, consciousness, and learning. Modeling such a cognitive cycle as a *cognitive atom* entails considering it as *indivisible*, as opposed to dividing it into a sequence of separate tasks. To our understanding, LIDA acknowledges this indivisibility by implementing cognitive cycles through asynchronous distributed processes. Since LIDA's design imposes no synchronicity on the distributed processes that generate the cognitive cycle, it is indeed not divided into a predefined sequence. In our view, this approach embodies CCCA in which “atom” denotes a sense of indivisibility. We instead introduce a more radical view that uses “atom” as a primitive notion for designing the model.

Our Radical Interactionism conceptual commitment invites designers of cognitive models to consider the notion of *sensorimotor interaction* as a primitive, instead of perception and action. This is analogous to a mathematical system’s primitives, which are used in axioms and theorems to define more complex structures, but which are themselves undefined within the system. Traditional cognitive models take perceptions and actions as primitive notions, and derive the notion of sensorimotor interaction from them. The RI conceptual commitment recommends doing the opposite. Although the CCCA commitment takes a first step in the RI direction by considering cognitive cycles as indivisible, Franklin et al. still define a cognitive cycle using the primitive notions of observation and action. The RI commitment eliminates these as primitives and instead frames each cognitive cycle as a sensorimotor interaction.

Intuitively, a sensorimotor interaction fits Franklin et al.’s description of a cognitive cycle: “Each cycle constitutes a unit of sensing, attending and acting. A cognitive cycle can be thought of as a moment of cognition, a cognitive moment” (p4). Within constructivist epistemology, sensorimotor interactions can be viewed as representing a Piagetian (1955) *sensorimotor scheme*, from which the subject constructs knowledge of reality. At a philosophical level, this view relates to phenomenology (e.g., Dreyfus, 2007), which argues that knowledge of the self and of the world derives from regularities in phenomenological experience. Accordingly, a sensorimotor interaction may be understood as a *chunk of phenomenological experience*, making the stream of sensorimotor interactions the primitive material from which the agent constructs all of its knowledge.

2. RI Impact on cognitive modeling

As an example to illustrate the RI commitment, we examine how it would impact the LIDA model. It would imply modifying the left-side part of Figure 1 in Franklin et al.’s paper as shown in our Figure 1 below. Since perception and action do not exist in RI, the *Sensory Memory* and the *Motor Plan Execution* modules would be removed, as well as the *Sensory Stimulus*, *Actuator Execution*, and *Dorsal Stream* connections. Instead, the *Sensory Motor Memory* module would

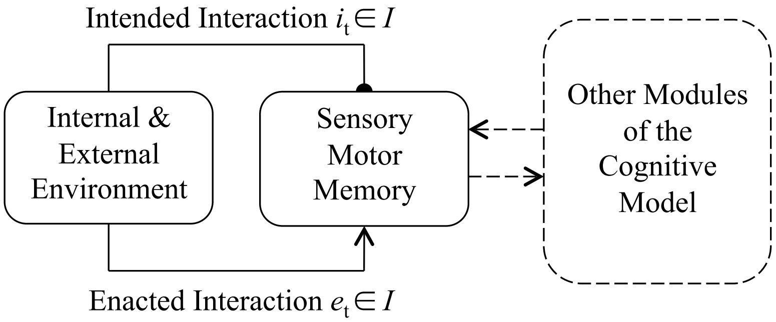
Intended Interaction $i_t \in I$
Internal & External Environment
Sensory Motor Memory
Other Modules of the Cognitive Model
Enacted Interaction $e_t \in I$

Figure 1. A diagrammatic adaptation of LIDA to Radical Interactionism. *Sensory Motor Memory* directly interacts with *Internal & External Environment* through the *Intended Interaction* i_t and the *Enacted Interaction* e_t. The *Other Modules of the Cognitive Model* box encompasses the remaining modules of LIDA. At time *t*, the agent chooses the sensorimotor interaction i_t that it intends to enact from among the set of sensorimotor interactions *I*. The attempt to enact i_t may change the environment. The agent then receives the enacted sensorimotor interaction e_t.

directly connect to the *Internal & External Environment* module through the *Intended Interaction* and the *Enacted Interaction* connections. As a result of these changes, *stimuli* and *actions* would be eliminated from the architecture and replaced by a single type of primitive object: sensorimotor interactions.

The algorithm begins with a predefined set of sensorimotor interactions *I*, called *primitive interactions*. At a given time *t*, the agent chooses a primitive interaction i_t that it intends to enact, from among *I*. The agent ignores this enaction's meaning; that is, the agent has no rules that would exploit knowledge of how the designer programmed the primitive interactions through actuator movements and sensory feedback (such as: "if a specific interaction was enacted then perform a specific computation"). As a response from the tentative enaction of i_t, the agent receives the *enacted interaction* e_t, which may differ from i_t. The enacted interaction is the only data available to the agent that carries some information about the external world, but the agent ignores the meaning of this information. As an example, the primitive interaction i_t may correspond to actively feeling (through touching) an object in front of the agent, involving both a movement and a sensory feedback. The tentative enaction of i_t may indeed result in feeling an object, in which case $e_t = i_t$, or may result in feeling nothing, if there is no object in front of the agent, in which case the enacted interaction e_t corresponds to a different interaction: moving while feeling nothing. The agent constructs knowledge about its environment and organizes its behavior through regularities observed in the sequences of enacted interactions.

Within the RI commitment, LIDA would not have to construct *action schemes* defined "as an action together with its context and expected result" (Franklin et al., p5, citing Drescher, 1991). Procedural and episodic memories would instead directly process sensorimotor interactions. The learning mechanism would construct higher-level sensorimotor interactions, called *composite interactions*, as pairs of a *context interaction* and an *intention interaction*: $i_{\text{composite}} = \langle i_{\text{context}}, i_{\text{intention}} \rangle$. Such a learning mechanism would be bottom-up, and would generate a hierarchy of composite interactions where each level would contain lower-level composite interactions, except the bottom level, which would be made of primitive interactions. At any time, the agent would represent its current situation as a set of interactions, called interactional context. The LIDA behavior selection mechanism would be modified to activate composite interactions $i_{\text{composite}}$ whose context interaction i_{context} belongs to the interactional context. The intention interactions $i_{\text{intention}}$ of the activated composite interactions $i_{\text{composite}}$ would compete to be the one selected as the next intended interaction. Since the intended interaction can be a primitive interaction or a previously learned composite interaction, this mechanism allows asynchrony between the architecture's behavior-selection mechanism and the primitive interactional mechanism, as illustrated in Figure 2.

The coupling between the cognitive architecture and the environment (represented by the *Internal & External Environment Constructed at Time t* in Figure 2) evolves as the agent learns increasingly sophisticated patterns of behaviors, while simultaneously representing the world in terms of increasingly sophisticated affordances. We refer the reader to other publications for a more thorough description of this mechanism (Georgeon & Ritter, 2012), its implications in a cognitive architecture (Georgeon, Marshall, and Manzotti, 2013), and demonstrations in which primitive interactions consist of active feeling (e.g., Georgeon and Marshall, 2013) and rudimentary active vision (Georgeon, Cohen, and Cordier, 2011).

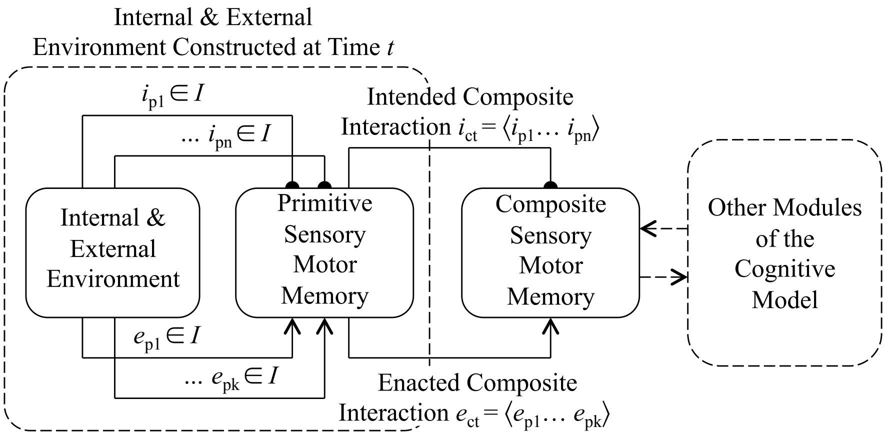
Internal & External Environment Constructed at Time t
$i_{\mathrm{p1}} \in I$
... $i_{\mathrm{pn}} \in I$
Intended Composite Interaction $i_{\mathrm{ct}} = \langle i_{\mathrm{p1}} \dots i_{\mathrm{pn}} \rangle$
Internal & External Environment
Primitive Sensory Motor Memory
Composite Sensory Motor Memory
Other Modules of the Cognitive Model
$e_{\mathrm{p1}} \in I$
... $e_{\mathrm{pk}} \in I$
Enacted Composite Interaction $e_{\mathrm{ct}} = \langle e_{\mathrm{p1}} \dots e_{\mathrm{pk}} \rangle$

Figure 2. Levels of interaction in Radical Interactionism. Over time, the agent constructs composite interactions, which implement a series of primitive interactions. At time t, the behavior selection mechanism may select a previously constructed composite interaction $i_{\mathrm{ct}} \approx \langle i_{\mathrm{p1}} \dots i_{\mathrm{pn}} \rangle$ to try to enact. The tentative enaction of i_{ct} is delegated to the *Primitive Sensory Motor Memory*, which controls the enaction of the sequence of the n primitive interactions i_{p1} to i_{pn}, and returns the enacted composite interaction e_{ct} to the *Composite Sensory Motor Memory*. The rest of the cognitive architecture sees the primitive loop as an *Internal & External Environment Constructed at Time t*, with which it interacts through higher-level interactions. Since the environment that the cognitive architecture interacts with evolves as the agent develops, the agent can recursively learn increasingly complex behaviors.

3. Conclusion

Rather than offering a method to address the traditional perception-action problem, Radical Interactionism presents a profoundly different formulation of this problem. Indeed, many studies define the perception-action problem as learning a mapping between a predefined action space and a predefined observation space. For example, these include studies in reinforcement learning (e.g., Sutton & Barto, 1998), means-end analysis (e.g., Drescher, 1991), and robotics studies (e.g., Pierce & Kuipers 1997). In contrast, the RI approach begins with a predefined interactional space, and focuses on the problem of generating learning effects that are considered important in AGI agents, namely self-programming (e.g., Thórisson, Nivel, Sanz, and Wang, 2013) and *constitutive autonomy*, which Froese and Ziemke (2009) define as an agent's ability to "self-constitute its identity", which they argue is a prerequisite for autonomous sense-making. RI agents realize self-programming because they learn increasingly long sequences of interactions and re-enact these in appropriate contexts. They have constitutive autonomy because their coupling with the environment evolves through their individual experience of interaction.

Moreover, RI allows implementing a type of self-motivation called *interactional motivation* (Georgeon, Marshall, and Gay, 2012). To implement interactional motivation, a designer attaches a predefined intrinsic valence to some primitive interactions, and biases the behavior selection mechanism to select sequences of interactions that have the highest total valence. Interactional motivation provides a way to specify inborn preferences (some primitive interactions that the agent innately likes or dislikes) but does not specify the agent's goal; the agent thus engages in open-ended learning to find its own way to fulfill its inborn preferences of interaction. This view implements Glasersfeld's (1984, p29) idea that goals "arise for no other reason than this: a cognitive organism evaluates its experiences, and because it evaluates them, it tends to repeat ones and avoid others".

At the theoretical level, RI relates to Glasersfeld's radical constructivism, whose "radical difference [from traditional conceptualizations] concerns the relation of knowledge and reality. Whereas in the traditional view of epistemology, as well as of cognitive psychology, that relation is always seen as a more or less picture-like (iconic) correspondence or match, radical constructivism sees it as an adaptation in the functional sense" (Glasersfeld, 1984, p20). Accordingly, RI does not implement a "picture-like" relation in the input received by the model from the environment. In contrast with observations in traditional models, the input received by an RI model from the environment only consists in the enacted interaction, which does not directly represent the environment. Because the input does not directly represent the environment, RI agents can be designed without ontological assumptions about the environment, which is critical for agents that perform open-ended learning in the real world.

Acknowledgements

This work was supported by the French Agence Nationale de la Recherche (ANR) contract ANR-10-PDOC-007-01, and NRL. The views and opinions contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of NRL or the DoD.

References

Drescher, G. L. 1991. *Made-up minds, a constructivist approach to artificial intelligence*. Cambridge, MA: MIT Press.

Dreyfus, H. 2007. Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology*. 20(2): 247-268.

Froese, T. and Ziemke, T. 2009. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*. 173(3-4): 466–500.

Georgeon, O.; Cohen, M.; and Cordier, A. 2011. A model and simulation of early-stage vision as a developmental sensorimotor process. In *proceedings of Artificial Intelligence Application and Innovations (AIAI'2011)*, 11-16. Corfu, Greece.

Georgeon, O. and Ritter, F. 2012. An intrinsically-motivated schema mechanism to model and simulate emergent cognition. *Cognitive Systems Research*. 15-16: 73-92.

Georgeon, O.; Marshall, J.; and Gay, S. 2012. Interactional motivation in artificial systems: between extrinsic and intrinsic motivation. In *proceedings of the Second International Conference on Development and Learning, and on Epigenetic Robotics (EPIROB'2012)*, 1-2. San Diego, CA.

Georgeon, O. and Marshall, J. 2013. Demonstrating sensemaking emergence in artificial agents: A method and an example. *International Journal of Machine Consciousness.* 5(2): 131-144.

Georgeon, O.; Marshall, J.; and Manzotti, R. 2013. ECA: An enactivist cognitive architecture based on sensorimotor modeling. *Biologically Inspired Cognitive Architectures*. 6: 46-57.

Glasersfeld, E. V. 1984. An introduction to radical constructivism. In *The invented reality* ed. P. Watzlawick. 17–40. New York, NY: Norton.

Piaget, J. 1955. *The construction of reality in the child*. London: Routledge and Kegan Paul.

Pierce, D. and Kuipers, B. 1997. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence*. 92: 169-227.

Sutton, R. and Barto, A. 1998. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Thórisson, K.; Nivel, E.; Sanz, R.; and Wang, P. 2013. Approaches and Assumptions of Self-Programming in Achieving Artificial General Intelligence. *Journal of Artificial General Intelligence*. 3(3): 1–10.

Commitments of the Soar Cognitive Architecture

John E. Laird LAIRD@UMICH.EDU
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109-2121

Abstract

I describe many of the important commitments in Soar. Although there are significant overlaps with the conceptual commitments for LIDA, there are additional commitments in Soar to functional requirements for artificial general intelligence (AGI) that distinguish it from LIDA. Some of these requirements include developing a full computational implementation, achieving real-time performance, implementing complex cognitive capabilities, and building agents that use those capabilities for multiple real world tasks.

1. Introduction

The Soar cognitive architecture (Laird, 2012) has been under development for over thirty years. At an abstract level, Soar shares many of the conceptual commitments described by Franklin et al. (2013) for LIDA. However, our work on Soar differs from LIDA in that although it is biologically inspired, we are focused on the goal of artificial general intelligence. That goal leads to a commitment to a set of functional requirements that have shaped the structure of Soar, differentiating it from LIDA in the details of its conceptual commitments and implementation.

2. Commitment to AGI

To understand a cognitive architecture, you must understand the goals of its developers. My goal for Soar is to develop a cognitive architecture that supports artificial general intelligence (AGI). Other possible goals for cognitive architecture research include modeling the details of the human brain and modeling the details of the human mind, as in ACT-R (Anderson, 2007), EPIC (Kieras & Meyer, 1997), LEABRA (O'Reilly et al., 2012), and SPAUN (Eliasmith, 2012). At one time, Soar was used to pursue a unified theory of human cognition (Newell, 1990); however now AGI is my primary goal. I draw inspiration and guidance from the function and structure of the human mind and brain in designing and developing Soar, but it is beyond my capabilities to also meet all the constraints that arise from psychology and neuroscience. Furthermore, the research projects listed above are making significantly more progress than I ever could. Franklin et al. (2103) aspire to create an architecture that models the human mind and brain while concurrently achieving AGI. I am skeptical as to whether that is possible, and I am very interested to see how their work progresses, especially in comparison to these other efforts.

3. Commitment to Functional Requirements

A second, related, commitment of mine is to respond to the functional requirements of general intelligence. Functional requirements for general intelligence arise from characteristics of (1) the agent's environment, (2) the tasks it must perform, and (3) the agent's embodiment, including its perceptual and motor systems and computational substrate. Until we have a clear understanding of these requirements, it is difficult to know whether we are making progress much less compare and contrast alternative approaches. In Laird (2012), I attempt to take a first step by deriving characteristics of environments, tasks, and agents that lead to thirteen functional requirements for AGI agents. Each requirement implies a commitment to functionality that an architecture must support. Below are some of those commitments that distinguish our approach from LIDA's.

1. **Computational Implementation**. Cognitive architectures are complex beasts and without computer implementations it is impossible to predict their behavior, much less make claims about their ability to support general intelligence. To quote Herb Simon: "In the computer field, the moment of truth is a running program; all else is prophecy." To make matters worse, in an AGI agent, there are always unexpected dynamic interactions that arise from the multitude of components that make up the underlying architecture. In Soar, the conceptual development of our theories is intertwined with building, testing, and evaluating them in real-world agents. Given the uneven state of implementation of the LIDA cognitive theory, it is difficult to evaluate it as a potential architecture for general intelligence.
2. **Real-time Performance**. An AGI must respond to the dynamics of the real world in real time. In Soar, we are committed to (obsessed with?) real-time execution. A cognitive architecture must not only have the right functionality, but it must be able to deploy that functionality in real time, and maintain real-time performance as it scales to large bodies of knowledge. In Soar, this commitment shapes its design and has paid off: Soar runs 10-1,000 times faster than real time on standard hardware, even with millions of rules and memory elements, running for days of simulated real time. Ignoring this constraint in an architecture (such as LIDA) raises questions about its ability to scale to real-world problems.
3. **Complex Cognitive Capabilities**. Soar has also been shaped by our commitment to the complex cognitive capabilities that distinguish human-level intelligence such as relational representations, planning, hierarchical task decomposition, meta-reasoning, mental imagery,

natural language processing, and many forms of learning. Some of these are shared by LIDA, but so far the emphasis has been on much more primitive aspects of cognition.

4. **Complex Agents**. Franklin et al. state: “It is at least difficult, and likely impossible, to compare the efficacy of such architectures without building working AGI agents which, currently, seems not possible.” Although we are cannot build complete AGI agents today, we can build and evaluate agents that are on the path to AGI—agents that use complex cognitive capabilities to solve many challenging tasks in real-world environments. Only by building complex agents do we learn how our architectures succeed and fail at meeting the requirements of an AGI. Throughout Soar’s development we have pushed the envelope of agent complexity and capabilities. Currently, we are developing Rosie (Mohan et al. 2013), a Soar agent that uses natural language instruction to learn new adjectives, nouns, verbs, and prepositions, as well as new tasks in a fully embodied robot. Rosie learns new tasks such as Tower of Hanoi or Tic-Tac-Toe from scratch using interactive situated instruction, and she then executes those tasks using a robot arm and a Kinect sensor. Rosie uses all of the major components of Soar, and she allows us to pursue one of the core questions of artificial general intelligence: how can an agent *learn* to represent and pursue novel tasks without being programmed to perform those tasks ahead of time?

4. Commitment to Architectural Mechanisms

Our third set of commitments is to the structure of the computational architectural mechanisms that define Soar. Many of these overlap with the conceptual commitments in LIDA, and there are many similarities in the architectural mechanisms used in Soar to LIDA (and ACT-R). There is a central working memory with decay (similar to the Global Workspace), there is a long-term procedural memory that controls behavior, there are additional long-term memories that contain semantic and episodic knowledge that can be retrieved into working memory, and there are multiple online learning mechanisms. There is also a primitive cognitive cycle. Explicitly listing (and comparing) these commitments is important for progress in the field, for although there are many similarities in our commitments, there are important differences in their details. For example, in Soar, ACT-R, EPIC, and SPAUN, the cognitive cycle is “leaner” with significantly less processing than is in LIDA’s (modeled at approximately 50 ms. instead of 200-300 ms.). Given the multitude of published models that empirically support the commitment to the faster cycle time, this difference raises many research questions for LIDA.

Although listing and comparing commitments is a good start, we still need to agree on the functional requirements for AGI, and we need to develop a suite of tasks that embody those requirements. The tasks must require using those complex cognitive capabilities that distinguish human intelligence. One of the most important is *taskability*, where the agent *learns* tasks on its own, without the shortcuts and hacks that are all too often inherent with human programming.

References

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

Eliasmith, C. 2013. *How to Build a Brain.* Oxford: Oxford University Press.

Franklin, S., Strain, S., McCall, R., and Baars, B. 2013. Conceptual Commitments of the LIDA Model of Cognition, *Journal of Artificial General Intelligence*, 4(2), 1-22.

Kieras, D. E., and Meyer, D. E. 1997. An Overview of the EPIC Architecture for Cognition and Performance with Application to Human-Computer Interaction. *Human-Computer Interaction*, 12, 391-438.

Laird, J.E. 2012. *The Soar Cognitive Architecture,* Cambridge, MA: MIT Press.

Mohan, S., Kirk, J., and Laird, J. 2013 A Computational Model of Situated Task Learning with Interactive Instruction. *Proceedings of the 12th International Conference on Cognitive Modeling*. Ottawa, Canada.

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

O'Reilly, R.C., Hazy, T.E. and Herd, S.A. (2012). The Leabra cognitive architecture: How to play 20 principles with nature and win! S. Chipman (Ed.) *Oxford Handbook of Cognitive Science*, Oxford: Oxford University Press.

Rosenbloom, P. 2013. The Sigma cognitive architecture and system. *AISB Quarterly*, 136, 4-13.

Conceptual Commitments of AGI Projects

Pei Wang PEI.WANG@TEMPLE.EDU
Department of Computer and Information Sciences, Temple University
Philadelphia, PA 19122, USA

1. The Issue of Conceptual Commitments

The target article by Franklin *et al*. (in the current volume) addresses a crucial issue in Artificial General Intelligence (AGI), that is, the “conceptual commitments” of an AGI project.

Every scientific research project is based on some statements that establish the assumptions and destination of the research, as well as the criteria for the evaluation of progress. In other fields, such a foundation is usually provided either by a commonly accepted theory to be followed, or by some commonly admitted observations to be explained. The situation in AGI is unusual because the field has neither of the two. AGI is not based on a theory that is accepted by the majority of the researchers in the field. Currently, many theories are used by the researchers, and most of them are adopted from another field, such as computer science, mathematics, psychology, neuroscience, etc., and their applicability to the AGI problem has not been fully established yet. On the other hand, what observations are relevant to AGI is also a controversial issue. In a very rough sense, all AGI projects attempt to reproduce the “intelligence” as displayed by the human mind, but since the human mind can be described at different levels and scales, what descriptions are really related to AGI is not self-evident.

Consequently, every AGI researcher makes some conceptual commitments on the objective (*What should be done?)* and strategy (*How to do it?*) of AGI, according to their considerations. Such commitments usually cannot be proved formally (since they decide which formal model is applicable) or verified empirically (since they decide which types of data are relevant). Even so, it does not mean every commitment is equally justifiable, and improper commitments will lead the research to wrong directions. It is also possible for a project to be based on implicitly inconsistent commitments, which will cause serious troubles that are not easy to recognize at the early stage of the project. Very often a bad assumption causes bigger problems than mistakes within a formal model or a computer system, since it tends to become invisible to the researcher. This issue has special importance to AGI at the current moment, because the conceptual commitments behind an AGI project are often not clearly spelt out, which cause misunderstandings and confusions.

For the above reasons, I highly appreciate the authors of the target article for clarifying the conceptual commitments behind the LIDA project. I share some of their commitments, but not all of them. In the following, I will focus on the two major topics mentioned previously.

2. Commitments on Research Objective

In a broad sense, all AGI and AI projects take the human mind as the source of inspiration. Therefore, the difference in research objectives is not really "humanly vs. rationally", as suggested in Russell and Norvig (2010), since our notion of rationality comes from nowhere but the human mind, and, on the other hand, few AI researcher has proposed to duplicate a human feature without any rationalization about why it is needed by a computer, which is different from a human being in many fundamental aspects. For example, the authors of the target article make it clear that "LIDA is a biologically inspired model, which incorporates the useful, functional aspects of the brain into the model". Therefore, the real decisions for an AGI project is *where* to be similar to the human mind and *why* this similarity is desired. As argued in Wang (2008), there are multiple alternatives for this decision, and each of them can produce results with theoretical and practical values, though they correspond to different research paradigms.

In this aspect, the commitments of LIDA, or the *working definition* of "intelligence" accepted in this project, is different from that of mine, which is at the foundation of the AGI project NARS (Wang, 2006). While LIDA attempts to duplicate the *human cognitive functions* as studied in cognitive psychology (and several other disciplines), NARS attempts to duplicate the *rational principles* as studied in logic and decision theory (and several other disciplines). Because of this difference in research objective, NARS does not accept some of LIDA's commitments, like "Biologically Inspired", "Non-linear Dynamics Bridge to Neuroscience", and "Theta Gamma Coupling from the Cognitive Cycle". After all, what NARS wants to do is to achieve *adaptation with insufficient knowledge and resources* in a computer system, which is not biological by nature. Therefore, unless there is a *functional necessity*, the system will not be designed "like a human" in a certain aspect. The target article acknowledges the possibility of a non-biologically-inspired AGI, though does not go further to compare these two alternatives.

On the other hand, many of LIDA's commitments are indeed necessary for intelligence in general, such as "Embodied (Situated) Cognition", "Comprehensive Decay of Representations and Memory", and "Feelings as Motivators and Modulators of Learning". In NARS, similar commitments are made, though they are not introduced individually, but all derived from the same principle of *adaptation with insufficient knowledge and resources*, without using any biological terminology. For example, embodiment is not interpreted as "to have a human-like sensorimotor mechanism", but "to behave and to adapt according to the system's *experience* (i.e., the interaction between the hosting device and the environment)"; forgetting and feeling are judged to be necessary mainly because of the shortage of resources (Wang, 2006).

3. Commitments on Research Strategy

LIDA's conceptual commitment on "Systems-Level Modeling" effectively distinguishes AGI models from mainstream AI models. An ordinary AI project only addresses certain processes, functions, or capabilities of the human mind, mainly for the sake of simplicity and feasibility. Though there are calls for combining AI techniques (Brachman, 2005), it is still widely assumed that AI should follow the common *divide-and-conquer* methodology of computer science.

Most AGI researchers disagree with this "division-then-combination" strategy, because each of the traditional AI techniques is designed under very different assumptions and requirements, so they usually cannot be simply bundled together to form a coherent system. The design of a general-purpose system has its own special issues, which are very different from the issues in the design of special-purpose systems. This is a major reason for AGI to become an independent research field.

Though the AGI community shares the commitment of building a *complete* intelligent system, there are still two different ways to achieve it:

- The *integrated* approach: to design the system as an architecture with multiple modules, each of which is specialized for a cognitive function (Newell, 1990).
- The *unified* approach: to depend on a few core principles, with extensions and augments here or there, to realize multiple cognitive functions (Hutter, 2005; Wang, 2006).

Though the projects following each of the two approaches are very different from each other in details, it is still meaningful to compare the two without touching those details.

LIDA follows the integrated approach, which is intuitively appealing, given its divide-and-conquer style plus the integrity provided by the architecture. Even so, there are still some major challenges faced by every project following this approach.

The first issue is the criterion of modularization. So far, the proposed cognitive architectures partition the overall function in very different ways, even though there are some overlap. Even the consensus are mostly based on the existing practices in cognitive psychology and mainstream AI, where the trouble caused by improper divisions of functions is exactly what AGI attempts to avoid. In such an architecture, the introduction of a module is usually justified by the evidence on the existence of such a process or mechanism in the human mind, but even if that is indeed the case, these evidence usually does not indicate that this process or mechanism can be modeled independently of the other processes or mechanisms. Here an important point is that though it is often valid for the psychologists to focus on a certain mechanism X of the human mind, it is not equally valid for the AI researchers to do the same. It is the case because in AI it means that X can be *realized* as a stand-alone process without the other co-existence mechanisms; while in psychology it is only required for X to be *described* separately, without talking about the other coexisting mechanisms. The former is a much stronger claim than the latter.

For example, LIDA consists of several separate memory modules for different types of knowledge, while a unified system (like NARS) may only have a single memory that maintains different types of knowledge altogether. The target article does not argue for the necessity or advantages of using multiple memory modules, compared to a single memory that handles all these types of knowledge uniformly. The introduction of other modules need similar justifications — among so many cognitive functions, why these are selected as modules?

A related second issue is the coordination of the modules. In principle, it can often be argued that in a given architecture almost every module should be related directly to every other module. For example, why in LIDA the “Attention Codelets” only works on the “Current Situational Model”, but not on the memory modules and the “Global Workspace”? Attention is arguably involved in all types of storage mechanism in an AGI, but to consider all possible relations among a large number of modules is clearly infeasible for engineering considerations.

The above issues do not exist in the unified approach, though there are other challenges, especially on how to provide a relative simple explanation for phenomena as complicated as intelligence. Since that topic is beyond this discussion, it will not be addressed here.

References

Brachman, R. J. 2006. (AA)AI — more than the sum of its parts (AAAI Presidential Address). *AI Magazine* 27(4):19-34.

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin: Springer.

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

Russell, S.; and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach, 3rd Edition*. Upper Saddle River, NJ: Prentice Hall.

Wang, P. 2006. *Rigid Flexibility: The Logic of Intelligence*. Dordrecht: Springer.

Wang, P. 2008. What do you mean by “AI”? In *Proceedings of the First Conference on Artificial General Intelligence*, 362-373. Amsterdam: IOS Press.

Will (dis)Embodied LIDA Agents be Socially Interactive?

Travis J. Wiltshire TWILTSHI@IST.UCF.EDU
Emilio J. C. Lobato ELOBATO@ IST.UCF.EDU
Florian G. Jentsch FLORIAN.JENTSCH@UCF.EDU
Stephen M. Fiore SFIORE@ IST.UCF.EDU
University of Central Florida
3100 Technology Parkway
Orlando, FL 32826, USA

1. Introduction

The target article by Franklin, Strain, McCall, and Baars (2013) provides a useful way forward in achieving the goals of artificial general intelligence (AGI). They explicate the conceptual commitments of the Learning Intelligent Distribution Agent (LIDA) model through specification of the importance of each commitment, not only to this model, but also to the goal of AGI.

Proponents of other cognitive architectures should engage in similar efforts to collaboratively approach the goal of AGI. While we view the overall LIDA approach as beneficial, and commend the authors for their extensive interdisciplinary grounding of the model in experimental psychology, cognitive neuroscience and artificial intelligence, we raise a set of inter-related issues concerning what else, from our perspective, might be necessary to achieve AGI. Therefore, the aim of our commentary is to constructively highlight and expand these ideas. Our focus is on the role of embodiment for learning and for social interaction and how these factors may help to improve the LIDA model, and AGI architectures, more generally.

2. Committed to Disembodied Embodiment

One of the theoretical commitments cited in the target article is embodied and situated cognition. However, the authors do not "insist on a robotic implementation" (Franklin et al., p .8) of their model, which is surprising as one might expect at least some form of insistence on robotic implementation when committing to an embodied approach. How can one eschew amodal symbols in favor of perceptual symbols without some form of embodiment with which to interact with the environment? At issue is the extent to which AGI can be achieved in the absence of embodiment. Undoubtedly, any AGI will be a learning system, but will it also be interacting with the world in ways similar to, or in support of, human interaction? Therefore, is it possible for a disembodied system to learn from the environment without perturbing that environment and receiving feedback from such interactions? While the argument could be made that a disembodied agent may have some capacity to interact with the environment, it is likely that the perturbatory bandwidth of an AGI agent would be significantly greater when embodied (Dautenhahn, Ogden, & Quick, 2002). Embodiment is essential for the learning and interactive capabilities of an AGI agent because the morphology of the agent and, in turn, the agent's sensorimotor capacities, both constrain and prescribe the types of cognitive capacities of the agent (Anderson, 2003; Pfeifer, Lungarella, & Lida, 2007). Further, these determine what is meaningful and relevant to that agent (Froese & Ziemke, 2009). Therefore, it is difficult for us to envision that the LIDA cognitive architecture, or any other for that matter, would approach the goal of AGI—a human-level intelligence—when instantiated in a disembodied state. Given this, we wonder if AGI researchers should consider a commitment to *human-like* embodiment as a means to better meet their goals (cf., Vernon, 2010). However, we acknowledge that, as Chella (2012) noted, the issue of embodiment is complex. In fact, it is still often debated in cases of human and animal cognition. As such, we commend the authors for committing to perceptual symbols and in short, our main purpose for posing these questions is to prompt the consideration of the scalability of the LIDA cognitive architecture across a range of disembodied and embodied platforms.

3. Socially Interactive?

As noted above, an additional factor is how AGI can be achieved without also encompassing social interaction. We would like to consider the socially interactive capabilities available to any agent that operates under the framework of the LIDA model, whether it be embodied or disembodied. As Mavridris (2012) noted, any AGI agent should be capable of leveraging the opportunities for action and interaction afforded by both the physical and social environment. We have recently argued that this is especially important in cases where robots are required to function and be perceived by humans as effective teammates (Wiltshire, Barber, & Fiore, 2013; see also Goodrich & Yi, 2013). We think this is a vital consideration for the overarching goals of

AGI given that humans, to our current knowledge, possess the most sophisticated form of social intelligence (Emery, Clayton, & Frith, 2007). As such, any artificial cognitive system that humans develop will be fundamentally embedded in an information rich *social* environment with the inherent goal of engaging in adaptive interaction (Anderson, 2003; Dautenhahn et al., 2002).

We acknowledge that research has not yet supported the notion of dedicated brain regions to specific social cognitive mechanisms, but rather a widespread re-use of various non-social regions (Pryzembel, Smallwood, Pauen & Singer, 2012). However, given that the LIDA model is a model of mind, and not of brain, we think conceptualizations of LIDA would be improved if there was some articulation of the types of socially interactive capabilities their model may exhibit in its current instantiation. With the scope of the target article focusing on commitments of the model, is it safe to say that the authors are not committed to a socially interactive model of cognition? Indeed, prior work by the authors has articulated that "a theory of mind process would be necessary for an artificial general intelligence (AGI) agent" (Freidlander & Franklin, 2008, p. 1). But recent theory and research suggests that humans employ more than just a theory of mind process for social interaction. As such, below we review some of the latest theory and research regarding social cognitive mechanisms in order to pose the question as to whether the LIDA model provides such capabilities or query the authors if it is not necessary to do so.

One of the major distinctions in social cognition theory and research is the comparison between online and offline social cognition (Pryzembel et al., 2012). Online social cognition can be characterized as a bidirectional and reciprocal case of *actual social interaction* between two or more agents where the mental states and behaviors of one agent dynamically influence the mental states of the other, and in turn, that agent's behavior. Conversely, offline social cognition tends to be characteristic of cases where one agent acts as a passive observer of others. The latter has tended to be the focus of research in social and cognitive psychology for the past 30 years as well as neuroscience over the past decade (Pryzembel et al., 2012). We make this distinction because there is a growing body of empirical evidence that suggests that the processes employed in online versus offline social cognition are distinct (e.g., Tylén, Allen, Hunter, & Roepstorff, 2012). Likewise, Pezzulo (2012) makes this distinction in terms of the types of mindreading tasks an observer versus an actor can employ in what he calls the "Interaction Engine", which specifies these mechanisms across a number of types of observation and interaction scenarios and computational means for instantiating these.

Extending upon this distinction, recent integrative accounts of social cognition argue that a number of distinct research findings in this field, when taken together, align with dual-process theories of cognition (e.g., Bohl & van den Bos, 2012). In such accounts, traditional approaches to theory of mind, often based on studies of offline social cognition, are aligned with Type 2 cognitive processes that are characterized as explicit, controlled, and stimulus independent. Conversely, theories of social cognition emphasizing social interaction, and thus online social cognition, are characterized by Type 1 processes that are implicit, automatic, and stimulus-dependent (e.g., Bohl & van den Bos, 2012). Further, in this context, social cognitive neuroscience supports a systems-level distinction as well as interdependency between the two general types of cognitive processes (Pryzembel et al., 2012; Satpute & Lieberman, 2006). The point here is to place an emphasis on this body of work in order to consider how to instantiate these mechanisms. We have recently begun to articulate a set of modeling recommendations that take into account these recent advances in social cognition with the goal being to instantiate Type 1 and Type 2 mechanisms in embodied robots. In the case of Type 1 mechanisms, this would allow for a more direct understanding of, and automatic interaction with, the social environment. Contrarily, in the case of Type 2 mechanisms, this would allow for more complex and deliberate forms of social cognition, such as simulative and inferential mechanisms, that allow for prediction and interpretation of complex social situations (Wiltshire et al., 2013).

Advancing the functional social capabilities of complex cognitive architectures additionally requires the assessment of how such architectures are perceived by humans, particularly in cases of aspiration towards the goal of AGI. It is not only important to ensure that the social signals conveyed by an AGI agent are perceived by humans as intended (Vinciarelli, Pantic, & Bourard, 2009). For AGI agents, it may also be necessary to assess the degree to which the agents are more generally perceived as conveying agency; that is, some degree of autonomy, situation understanding, and intentionality (Froese & Ziemke, 2009). Takayama (2012) has explicated a useful framework that demonstrates that people tend to differentially perceive the presence of agency along two dimensions. Much like the above dichotomies (online vs. offline social cognition), people can perceive an entity as having agency in-the-moment, upon reflection after an interaction, or even during both. We assert that it is important for an AGI agent to be perceived and treated as conveying agency, at least during an interaction, if not also upon later reflection. Research in human-computer interaction has consistently shown that people will automatically treat machines as if they convey agency, but not perceive them as such upon reflection (e.g., Nass & Moon, 2002). Further, human-robot interaction (HRI) research has shown that humans will sometimes explicitly consider robots as volitional or intentional agents, at least in certain contexts (e.g., Fiore et al., 2013; Kahn et al., 2012a, 2012b). Taken together, our point here is that there should not only be a conceptual commitment to social interactivity in AGI cognitive architectures, but also an evaluative commitment to the ways in which an agent operating under such architectures is perceived by humans (cf. Steinfeld et al., 2006).

4. Concluding Remarks

In short, we are not sure if the LIDA model, as currently instantiated, would allow for the types of interaction capabilities necessary for effective HRI. We have posed a number of questions about the commitment to disembodied embodiment, elaborated on recent advances in social cognition, and posited how to begin instantiating these in artificial agents as well as encouraged evaluating how an agent is perceived by humans given this additional social functionality. Our aim has been constructive in the hopes that the designers of the LIDA model can either make explicit the socially interactive capabilities of the LIDA model or at least begin to consider our recommendations as a way to advance the model in this capacity. Further, given that the LIDA model is not committed to an embodied implementation, the architecture will likely need to account for the distinction between Type 1 and Type 2 social cognitive processes as well the various types of mechanisms required for different interaction scenarios. In this vein, we recommend Pezzulo's (2012) Interaction Engine as a worthwhile "checklist" for assessing the interactive capabilities of the LIDA model. This can be assessed as a function of the type of embodiment the architecture is instantiated upon and the perturbatory bandwidth of that platform for the physical *and* social environment. Finally, we emphasize a commitment to a socially interactive cognitive architecture as we expect this will provide the basis for reciprocal intention understanding, effective communication, and shared task understanding (Pezzulo, 2012)—each, essential for effective human interaction with AGIs.

Acknowledgements

This work was partially supported by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0016. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the U.S. Government or the University of Central Florida. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Anderson, M. L. 2003. Embodied Cognition: A Field Guide. *Artificial Intelligence*, 149: 91-130.

Bohl, V.; and van den Bos, W. 2012. Toward an Integrative Account of Social Cognition: Marrying Theory of Mind and Interactionism to Study the Interplay of Type 1 and Type 2 Processes. *Frontiers in Human Neuroscience,* 6: 1-15.

Chella, A. 2012. Are Disembodied Agents Really Autonomous? *Journal of Artificial General Intelligence,* 3: 31-63.

Dautenhahn, L.; Ogden, B.; and Quick, T. 2002. From Embodied to Socially Embedded Agents: Implications for Interaction-Aware Robots. *Cognitive Systems Research,* 3: 397-428.

Emery, N. J.; Clayton, N. S.; and Frith, C. D. 2007. Introduction. Social Intelligence: From Brain to Culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362: 485-488.

Fiore, S.M.; Wiltshire, T.J.; Lobato, E.J.C.; Jentsch, F.G.; Huang, W.H.; and Axelrod, B. 2013. Towards Understanding Social Cues and Signals in Human-Robot Interaction: Effects of Robot Gaze and Proxemic Behavior. *Frontiers in Psychology: Cognitive Science,* 4: 1-15.

Franklin, S.; Strain, S.; McCall, R.; and Baars, B. 2013. Conceptual Commitments of the LIDA Model of Cognition. *Journal of Artificial General Intelligence*. 4: 1-22.

Froese, T.; and Ziemke, T. 2009. Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind. *Artificial Intelligence*, 173: 466-500.

Friedlandera, D.; and Franklin, S. 2008. LIDA and a Theory of Mind. In *Proceedings of the First Artificial General Intelligence Conference.* 171: 137-148. IOS Press.

Goodrich, M. A.; and Yi, D. 2013. Toward Task-Based Mental Models of Human-Robot Teaming: A Bayesian Approach. In *Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments*, 267-276. Springer Berlin Heidelberg.

Kahn, P. H.; Kanda, T.; Ishiguro, H.; Freir, N. G.; Severson, R. L.; Gill, B. T.; Ruckert, J. H.; and Shen, S. 2012a. "Robovie, you'll have to go into the closet now": Children's Social and Moral Relationships with a Humanoid Robot. *Developmental Psychology.* 48: 303-314.

Kahn, P. H.; Kanda, T.; Ishiguro, H.; Gill, B. T.; Ruckert, J. H.; Shen, S.; and Severson, R. L. 2012b. Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes? In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 33-40. Boston, MA. Association for Computing Machinery.

Mavridis, N. 2012. Autonomy, Isolation, and Collective Intelligence. *Journal of Artificial General Intelligence,* 3: 31-63.

Nass, C.; and Moon, Y. 2002. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues,* 56: 81-103.

Pezzulo, G. 2012. The "Interaction Engine": A Common Pragmatic Competence across Linguistic and Nonlinguistic Interactions. *IEEE Transactions on Autonomous Mental Development,* 4: 105-123.

Pfeifer, R.; Lungarella, M.; and Lida, F. 2007. Self-Organization, Embodiment, and Biologically Inspired Robotics. *Science,* 318: 1088-1093.

Przyrembel, M.; Smallwood, J.; Pauen, M.; and Singer, T. 2012. Illuminating the Dark Matter of Social Neuroscience: Considering the Problem of Social Interaction from Philosophical, Psychological, and Neuroscientific Perspectives. *Frontiers in Human Neuroscience*, 6: 1-15.

Satpute, A. B.; and Lieberman, M. D. 2006. Integrating Automatic and Controlled Processes into Neurocognitive Models of Social Cognition. *Brain Research,* 1079: 86-97.

Steinfeld, A.; Fong, T.; Kaber, D.; Lewis, M.; Scholtz, J.; Schultz, A.; and Goodrich, M. 2006. Common Metrics for Human-Robot Interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction.* 33-40. Salt Lake City, UT. Association for Computing Machinery.

Takayama, L. 2012. Perspectives on Agency: Interacting with and through Personal Robots. In *Human-Computer Interaction: The Agency Perspective* eds. M. Zacarias; and J. V. Oliveira, 195-214. Heidelberg, Germany: Springer.

Tylén, K.; Allen, M.; Hunter, B. K.; and Roepstorff, A. 2012. Interaction vs. Observation: Distinctive Modes of Social Cognition in Human Brain and Behavior? A Combined fMRI and Eye-Tracking Study. *Frontiers in Human Neuroscience*, 6: 1-11.

Vernon, D. 2010. Enaction as a Conceptual Framework for Developmental Cognitive Robotics. *Paladyn*, 1: 89-98.

Vinciarelli, A.; Pantic, M.; and Bourlard, H. 2009. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing*, 27: 1743-1759.

Wiltshire, T. J.; Barber, D.; and Fiore, S. M. 2013. Towards Modeling Social Cognitive Processes in Robots to Facilitate Human-Robot Teaming. In *Proceedings of the 57^{h} Annual Meeting of the Human Factors and Ergonomics Society.* 1278-1282. San Diego, CA.

Authors' Response to Commentaries

Steve Strain SFSTRAIN@MEMPHIS.EDU
Stan Franklin FRANKLIN@MEMPHIS.EDU
Fedex Institute of Technology 301
The University of Memphis
Memphis, TN 38152, USA

"History suggests that the road to a firm research consensus
is extraordinarily arduous." (Kuhn 1962, p. 15)

The twelve visually impaired men of Hindustan may yet succeed in understanding an elephant if they work together. Since each individual is incapable of perceiving the entire elephant at once, their ability to communicate determines whether all available evidence will be integrated into a unified model. Intelligence itself certainly evolved under selection for adaptive modes of social cooperation of this kind. What would an intelligent way of studying AGI look like?

We seek an answer to this question in the collective wisdom of the AGI community. In our target article, Franklin, Strain, McCall, and Baars, "Conceptual Commitments of the LIDA Model of Cognition" (*JAGI*, June 2013), we presented the tentative hypotheses of the LIDA Model of Cognition, and invited discussion among AGI researchers regarding which hypotheses are essential or useful to AGI. We are grateful to Angerer and Schneider (**A&S**); Chella (**AC**); Georgeon and Aha (**G&A**); Laird (**JL**); Wang (**PW**); and Wiltshire, Lobato, Jentsch, and Fiore (**WLJ&F**) for carefully reading the target article and for writing the commentaries that appear in this issue of *JAGI*. Here we review the salient points of the commentaries, respond to some of the insights and challenges presented in them, and seek to extend the discussion regarding desiderata for AGI.

A&S begin by observing the ambiguity regarding the term "AGI." They then proceed to caution against blithely labeling computational mechanisms and algorithms with cognitive monikers such as UNDERSTAND or GOAL. Rightly, they insist on careful justification of such labels in functional terms. They continue with remarks on the problem of comparing potential AGIs with a standard and with each other, and conclude with concerns about the modules in LIDA, claiming that these implicitly entail "a further conceptual commitment."

AC notes LIDA's strong commitment to consciousness, and emphasizes its importance as a key aspect of intelligence. **G&A** note that their Enactive Cognitive Architecture (ECA) shares LIDA's commitment to *Cognitive Cycles as Cognitive Atoms*, extending it into a more specific form which they term *Radical Interactionism*. In particular, they stress the atom's etymologically inherent notion of indivisibility and state that cognition is first and foremost a sensorimotor interaction while perception and action are derivative rather than primitive functions, in contrast with LIDA and other architectures.

JL presents a number of methodological commitments of the Soar cognitive architecture, including an insistence on implementations with real-time functionality that model cognitive processes unique to human-level intelligence, without substantially addressing the issue of conceptual commitments. **WLJ&F** raise concerns with LIDA's stance on embodiment, suggesting that it may be too weak, and question whether LIDA's commitments are sufficient for a model of social cognition, that latter being particularly relevant for human-robot interaction.

PW begins by raising a key issue for the growth of AGI as a science: how conceptual commitments shape research objectives and strategies. He goes on to critique the LIDA Model's commitment to biological inspiration, claiming that for AGI, any such inspiration must be justified by a specific functional necessity. He suggests for AGI the central goal of "achieving adaptation with insufficient knowledge and resources" (p. 40). He classifies LIDA as an integrated approach in an integrated-vs.-unified ontology, offering his NARS architecture as prototypical of the unified approach. He highlights several difficulties of the integrated approach, especially the risk for modularization to isolate processes that are cognitively interdependent.

1. Regarding Paradigms for Intelligence and Cognition

1.1 AGI as a Proto-science

Carnap (1966) lists three types of scientific concepts: classificatory, comparative, and quantitative. Classificatory concepts, he states, are freely defined, their only *a priori* requirement being consistency. On the other hand, comparative concepts must satisfy two criteria beyond convention: "they must apply to facts of nature, and they must conform to a logical structure of relations" (Carnap 1966, p. 58). Quantitative concepts must build on a pair of comparative concepts (eg "hotter than" and "the same temperature as") and must also define at least one method for measurement.

AGI's component disciplines define a broad set of quantitative concepts, including IQ, reaction time, spike rate, CPU speed, vocabulary size, and many others. However, neither this class of inherited concepts, nor that of concepts newly derived from AI applications--such as the total time used in analyzing chess positions, the final score of Jeopardy match against a human being, the length of a cognitive cycle, or the number of production rules applied per second of real time—contain any members currently proven to bear on AGI as a science.

The comparative concepts of "as well as a human" and "better than a human" are clearly relevant for AGI, but are too broad at present. Deep Blue and Watson can outperform humans at certain games, but the concepts that define their approach are likely insufficient--and could even prove unnecessary--for AGI. No other comparative concept is readily apparent. As **A&S** point out, this poses a significant challenge for the evaluation of potential AGI architectures. Thus, it is not yet possible to establish AGI as a science in the classical sense.

1.2 AGI and the Structure of Paradigms

PW wisely raises the issue of foundational principles that define research goals and objectives for a scientific discipline. Kuhn (1962) observes that a mature scientific paradigm provides "a strong network of commitments—conceptual, theoretical, instrumental, and methodological—[that] … relates normal science to problem-solving." (p. 42) Within the confines of such a consensus, research questions are assured of having a solution, and rules and steps are given for recognizing and achieving them. Scientific progress then becomes a matter of determining which facts are significant, increasing the match between theory and observation, and further articulating the paradigm. However, he points out that, "Often a paradigm developed for one set of phenomena is ambiguous in its application to other closely related ones" (Kuhn 1962, p. 29).

The latter point is especially pertinent for AGI. In *The Structure of Scientific Revolutions*, he noted as an exception fields "like biochemistry, [that] arose by division and recombination of specialties already matured" (1962, p. 15). Fifty years later, candidate paradigms proliferate like recombinant DNA as new evidence pushes mature sciences, such as those that compose AGI, to their limit. Rather than viewing these candidates as individuals vying for election to offices in a venerable institution, we see scientific investigation in the information era as an attractor landscape on the verge or in the midst of a massive phase transition. However germane his analysis of scientific history may be, it must not be assumed that Kuhn's conclusions will apply moving forward as well as they did in the 1960s. The present state of AGI (and of science in general, as we will explain below) is certainly not a period of "normal science," as Kuhn calls it, but rather, of "wild science." AGI may require a new order of paradigmatic organization. Even if one views this speculation as extreme, we urge open-mindedness and encourage rigorous examination of long-held assumptions. It is increasingly likely that some of the things we think we know for sure are not the case after all.

2. Regarding Biomimesis in AGI

While LIDA's conceptual commitment to biological cognition does indeed fall under the rubric of "biologically inspired," LIDA makes a more specific commitment to biological *consistency*. However, contrary to **JL's** claim, LIDA does *not* seek to model brain, but, as a general model of mind, it must model biological mind as well as artificial mind.

PW challenges the relevance of this goal to AGI. Deeming "the rational principles as studied in logic and decision theory" (p. 40) to be sufficient for AGI, he suggests that biosimilarity should occur not by direct design, but only by *a posteriori* analogy.

We point out that biological cognition offers the only extant exemplar of actual cognition. We believe the most practical way to achieve AGI is to find a general model of mind that can be instantiated in the biological or the artificial domain. Moreover, for the purpose of discovering such a general model, we do not limit ourselvespage to human cognition. The study and simulation of non-human cognition indeed lies outside the boundaries of traditional AGI, but we question whether an AGI can be successfully developed without a better understanding of mind in general.

2.1 The Importance of Biological Models of Cognition

The roundworm *Caenorhabditis elegans* has a nervous system composed of exactly 302 neurons and 7000 synapses, all genetically determined, that includes receptors for gustatory, olfactory, tactile, thermal, noxious, proprioceptive, and osmotic stimuli, and with which it can crawl, swim, eat, lay eggs, avoid obstacles, evade predators, and enter a sleep-like state (Greenspan 2007). No model currently provided by computational neuroscience can explain this. How many production rules would be required to model such compact behavioral richness, capable of survival in a pair of hostile environments as well as reproduction in the aqueous one? This system exhibits many orders of magnitude less complexity than the human nervous system. We consider *C. elegans* to represent a point along the evolution of biological cognition, a better understanding of which will—we argue—be highly relevant to the goals of AGI.

Some will complain, as does **JL**, that we conflate brain and mind; however, we do not. Rather, an understanding of the relation between a biological brain and its mind, at a high level at least, will be necessary to develop a general model of intelligence that can be instantiated on a non-biological substrate. How high a level of understanding is required remains an open question.

Sloman presents a model of mind-brain supervenience (2009) that plays a central role in our conceptual commitments. In brief, a supervenience relationship between two domains requires that any change in the state of the supervening system entail a change in the other system. Thus, mind supervenes on brain. However, mental state cannot be directly measured by physical means, which poses a dilemma for classical materialism. Sloman argues convincingly that bidirectional causal relationships between mind and brain exist and can be meaningfully defined. The relationship between thought and action is the essence of intelligence, be it human or non-human; thus we consider mind-brain supervenience crucial to the study of mind. This is very different from attempting to model mind and brain simultaneously. LIDA's theoretical model seeks to understand mind-brain supervenience at a high level in order to reproduce that relationship *in silico*. How does brain generate mind?

We doubt whether approaches to AGI that ignore or marginalize this crucial question will succeed. Many have invoked an analogy with the discovery of mechanical flight, claiming that as the Wright Brothers succeeded without modeling biological flight, so must biomimetic approaches to cognition be abandoned in order to achieve the optimal solution for AGI (eg Hinchey & Sterritt 2012, Pollack 1994, Russell & Norvig 2003). It was indeed possible to develop the airplane, the helicopter, and even the jet in advance of a thorough understanding of biological flight. But biological intelligence is much more similar to artificial intelligence than biological flight is to human flight, for the following reasons. First, biology never solved the problem of human flight. Avian and insect flight must support different functional requirements than mechanical aviation—including perching, hovering, darting, and soaring for food acquisition, predation, nest-building, surveillance, evading predators, social interaction, reproduction, and migration--and they utilize integrated strategies for lift, thrust, and control that do not scale with increasing load. When mechanical flight seeks similar functional requirements for a similar load, the biomimetic approach is extremely useful, as in the case of the Nano Hummingbird drone and others (Belski & Refosco 2012, Bumiller & Shanker 2011). Evolution is an exquisite designer.

Second, flight itself is intrinsically physical, and its principles are well-expressed in terms of aerodynamics, the principles of which govern living and non-living objects alike. On the other hand, cognition as we know it has evolved as an adaptation enabling successful management of increasingly complex behaviors focused on survival and reproduction, and in this way is intrinsically biological. The logic and rationality that, as **JL** and **PW** state, distinguish human cognition, are evolutionary endpoints; many aspects of biological intelligence take phylogenetic precedence over them.

AGI hopes to solve the exact problems already solved by biological cognition, namely, to generate adaptive behavior (including learning) on the basis of sensory input. Watching Deep Blue defeat Gary Kasparov at chess and Watson defeat Ken Jennings at Jeopardy are highly relevant benchmarks for machine intelligence, yet no one would think of having a distance race or a cargo carrying competition between an airplane and a bird. Fixed-wing flight enables the rapid transport of large loads across long distances, and can be adapted for military, meteorological, geographic, or other purposes, but biological flight allows, across a wide array of environmental conditions, its flexible and cost-effective application to the rapidly shifting task demands of an organism. Not only are the physical problems and goals quite different between the two cases, but for the latter, such behavioral richness and adaptability offer much of value to careful students of cognition.

2.2 Compartmentalization and Integration in Biological Cognition

A&S raise the concern that LIDA's modules may reinforce a possibly incorrect scheme of functional compartmentalization, inherited from cognitive psychology. They furthermore incorrectly state that said modules form an additional, implicit commitment of the LIDA Model. On the contrary, the modules themselves were deliberately left out of our list of commitments.

Rather than a commitment, they represent a working taxonomy of cognitive functions. They have evolved fluidly over LIDA's decade-long gestation, according to theoretical and practical demands. They do, however, entail a commitment to *some* kind of modular organization, although not of the type suggested by **PW**.

Ample evidence establishes functional compartmentalization in biological systems at multiple spatiotemporal scales and in diverse contexts (Alderson & Doyle 2010, Csete & Doyle 2002, Gerhart & Kirschner 2007). Unlike the compartmentalization typical of computer science and engineering, the biological variety maintains a controlled interdependence between component processes. This enhances behavioral robustness at the expense of added complexity. Computational modeling of this kind of adaptive complexity presents an enormous challenge, one that has not been adequately answered, but LIDA does accommodate it in principle, and our computational framework does attempt to address it by allowing interaction between modules in the form of codelets. The resulting implementational strategy in some ways resembles a multi-agent system (Peña et al. 2012 offers a review of software engineering principles that pertain in such a system).

PW's ontology of unified vs. integrated approaches, with its classification of LIDA as a "divide-and-conquer" strategy for AGI, fails to appreciate the nature of biological compartmentalization, which does *not* decompose a whole into independent processes. Biological cognition is both unified and integrated. In order to succeed in its overarching goal of understanding mind, LIDA must address this. Whether this is also necessary for AGI is another open question, but this possibility cannot be excluded unless and until a simpler AGI is in fact achieved.

2.3 Complexity in Biological and Other Systems

We further differ from **PW** regarding his interpretation of our commitment to "Systems-Level Modeling." Rather than implying the assembly of numerous AI components into a "patchwork quilt," our view of systems-level modeling considers dynamical as well as structural issues, a requirement of systems biology (Kitano 2002). While this indeed opens a *Pandora's Box* of theoretical and practical difficulties, it is desperation rather than grandiosity that compels us to do so. Closed systems and reductionistic models have proved enormously useful and insightful, and well deserve recognition, respect, and appropriate application, but they have repeatedly failed to address the problems posed by systems that exhibit real-world complexity.

Open systems do not obey conservation laws, and thereby operate far from the equilibria characteristic of closed physical systems. Attempts to break open systems into hierarchy of closed systems fail, for the reasons that **PW** suggests, and also reinforce theoretical errors, as indicated by **A&S**. One salient example followed the discovery of the genetic code. A number of gene sequences were discovered that, although preserved fairly well on an evolutionary scale, did not code for any known protein or transfer RNA. They were therefore widely assumed to have no function whatsoever, earning them the label "junk DNA." Recent work has proven this notion entirely incorrect (Brosius & Gould 1992; Pyle 2010; Willingham & Gingeras 2006), but as A&S mention, such misapprehensions are difficult to correct once they become widely disseminated. Furthermore, current evidence challenges the status of the triplet codon as "the" genetic code, suggesting that at least one distinct genetic code coexists with the widely known code for amino acid sequence (Narasimharao 2013; Peckham et al. 2007; Warnecke et al. 2008). While these observations seem far afield from AGI—a seeming that is itself an assumption that could mislead in the very same way--the tale they tell is compelling. Our field is much too complex for us to blithely exclude possibilities in the name of Ockham's razor and for the sake of intellectual comfort. We must face the fact and the inconvenience of our own present ignorance.

In truth, a kind of death knell may be sounding for classical theories in a number of fields. Lighthill, notorious for his role in precipitating the first AI winter, authored a 1986 paper admitting the failure of classical determinism in many systems governed by *Newtonian* laws. In the past century, many others have indicated similar shortcomings for reductionism and mechanistic modeling in a number of fields, including thermodynamics (Schrödinger 1944; von Bertalanffy 1950), neuroscience (Bullock 1993; Bullock et al. 2005; Freeman 2003), and biology (Maturana & Varela 1980; Polanyi 1968; Rosen 1970, 2000; von Uexküll 1926). In particular, Freeman, Maturana & Varela, and Rosen discuss the implications of this crisis for biological cognition. While extraordinarily daunting, it is extremely important that these dilemmas be brought into plain view rather than remaining on the sidelines. As we describe elsewhere (Strain et al. in preparation), LIDA's hypotheses regarding self-organizing dynamics and coordination of brain rhythms begin tying some of the major strands of this dizzying scene together.

2.4 Consciousness, Embodiment and Social Cognition

We thank **AC** for emphasizing LIDA's commitment to consciousness, and its importance for general intelligence. While we expressed this commitment implicitly in our commitment to Global Workspace Theory as a particular model of consciousness, it indeed merits more general mention. Consciousness has recognizable neural correlates (Baars 1988; Baars et al. 2013), predates human intelligence, and has likely played a significant role in the evolution of cognition (Edelman et al. 2005). We acknowledge that it may be possible to accomplish AGI without consciousness, but it certainly plays a role in biological intelligence, and it is indeed the core of the LIDA Model.

WLJ&F raise concerns regarding LIDA's commitment to embodiment, or lack thereof, in their view, which defines embodiment as a robotic implementation. Our view is more general: Embodiment can occur in a non-physical agent as long is the agent is structurally coupled with its environment (Franklin 1997). With this conception in mind, LIDA can accommodate robotic cognition, but views general intelligence as a more abstract entity.

WLJ&F also question whether LIDA can model social intelligence. While LIDA can in principle, this is an extremely challenging problem at the level of a general model of mind, and so far, little theoretical or empirical work has been done on this area by LIDA researchers. Biological mechanisms affecting social cognition range from the cellular (eg Gordon et al. 2013) to the behavioral and perhaps the cultural. The research of Marsh and colleagues (2006, 2009) resonates with LIDA's commitments to embodiment and self-organizing dynamics, as well as our interest in Gibsonian affordances. The work of Bohl and Van den Bos (2012) is new to us, and we are grateful for this reference.

G&A introduce the concept of Radical Interactionism (RI), which they describe as an extension of LIDA's *Cognitive Cycles as Cognitive Atoms* commitment in the context of their Enactive Cognitive Architecture (ECA). Inspired by the work of Varela et al. on embodiment (1991), ECA uses sensorimotor "interactions" in place of the traditional notion of perception and action as separate processes, a coupling described elsewhere in Gibsonian terms (Georgeon & Ritter 2012). Like LIDA, ECA employs a scheme mechanism inspired by Drescher, but modified to pair context and intention interactions rather than perceived contexts and actions. In this way, as well as through its commitments to embodiment and biological inspiration, it shares some similarity with LIDA, but it does not attempt to model biological cognition *per se* and does include a mechanism for consciousness.

3. Regarding Implementations

JL describes the methodological commitments of SOAR and critiques LIDA from this perspective. A key functional requirement of SOAR is its focus on computational implementation as the predominant mode of discourse for the field. Invoking Simon in support of this demo-or-die approach, he equates "the computer field" with AGI, and implied that AGI should be viewed as a sub-discipline of computer science.

This conception has long dominated mainstream AI, producing many remarkable implementations, but no AGI. In all fairness, this fact alone recommends the traditional perspective for serious re-examination. The ongoing stream of evidence has reinforced rather than weakened Dreyfus' critique of traditional AI (1972), leading to stronger and stronger challenges (eg see Lyon 2006).

Certainly AGI cannot exist without an implementation, but in view of the history of "AI Winters," evidence other than computational has long merited first class consideration. **JL** claims that SOAR's numerous implementations trace a path toward AGI, but similar claims were made by Minsky in 1968 (Dreyfus 2007). The truth of **JL's** claim remains to be seen. And it is indeed possible that pressure to implement at a certain rate could result in oversights that delay the discovery of AGI, as suggested by **A&S**.

The LIDA Model began its development after a successful implementation of human-like intelligence, the Intelligent Distribution Agent (IDA). IDA replicated the duties of a Naval detailer, namely, communicating with sailors via email regarding their reassignment preferences, consulting US Naval policy databases, and deciding which reassignment options to offer the sailors. Modeled after the actual thought process of detailers, IDA employed a domain-specific cognitive cycle and consciousness mechanism less complex but similar to those of LIDA. IDA's performance was judged by actual detailers and was deemed comparable to that of its human counterparts (Franklin et al. 1998, Baars & Franklin 2007). LIDA's commitment to biological, neuroscientific, and psychological modes of evidence has slowed its rate of growth relative to SOAR, but the implications of this for AGI are undetermined, given the field's status as a proto-science.

4. Conclusion

We greatly appreciate the time and effort expended by **A&S**, **AC**, **G&A**, **JL**, **PW**, and **WLJ&F** to provide these excellent commentaries. We consider the variety of viewpoints they demonstrate to be a very healthy finding given the inchoate state of our field. A strong consensus on AGI at this point in time would lack a sufficiently sound empirical basis: There is a surfeit of possibilities amidst a paucity of definitive empirical results. However, it seems both appropriate and constructive for each research group to develop its own consensus, and to engage in open debate characterized by cooperation between theoretically aligned groups, and healthy skepticism otherwise. This increases the probability that at least one AGI group is indeed on the actual path towards AGI. We look forward to the continuing discussion.

Acknowledgements

We would like to thank Sean Kugele for valuable insights and suggestions regarding the arguments presented in our response.

References

Alderson, D. L. and Doyle, J. D. 2010. Contrasting Views of Complexity and Their Implications For Network-Centric Infrastructures. *IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans*. 40(4):839-852.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. J. and Franklin, S. 2007. An Architectural Model of Conscious and Unconscious Brain Functions: Global Workspace Theory and IDA. *Neural Networks*. 20:955-961.

Baars, B. J.; Franklin, S.; and Ramsoy, T. Z. 2013. Global Workspace Dynamics: Cortical "Binding and Propagation" Enables Conscious Contents. *Frontiers in Psychology*. 4:200.

Belski, E. and Refosco, M. 2012. Emulation of Natural Flight with Biomimetic Wings. University of Pittsburgh, Swanson School of Engineering. Mechanical Engineering Section A4, March 1, 2012. http://136.142.82.187/eng12/Author/data/2040.docx.

Bohl, V; and van den Bos, W. 2012. Toward an Integrative Account of Social Cognition: Marrying Theory of Mind and Interactionism to Study the Interplay of Type 1 and Type 2 Processes. *Frontiers in Human Neuroscience*. 6:1-15.

Brosius, J. and Gould, S. J. 1992. On "Genomenclature": A Comprehensive (and Respectful) Taxonomy for Pseudogenes and Other "Junk DNA." *PNAS*. 89:10706-10710.

Bullock, T. H. 1993. *How Do Brains Work? Papers of a Comparative Neurophysiologist*. Boston: Birkhäusser.

Bullock, T. H.; Bennett, M. V. L.; Johnston, D.; Josephson, R.; Marder, E.; and Fields, R. D. 2005. The Neuron Doctrine, Redux. *Science*. 310(5749):791-793.

Bumiller, E. and Shanker, T. 2011. War Evolves with Drones, Some Tiny as Bugs. *The New York Times*. June 19, 2011.

Carnap, R. 1966. *An Introduction to the Philosophy of Science*. New York: Dover Publications.

Csete, M. E. and Doyle, J. C. 2002. Reverse Engineering of Biological Complexity. *Science*. 295:1664-1668.

Dreyfus, H. 1972. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row.

Dreyfus, H. 2007. Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology*. 20(2):247-268.

Edelman, D. B.; Baars, B. J.; and Seth, A. K. 2005. Identifying Hallmarks of Consciousness in Non-mammalian Species. *Consciousness and Cognition*. 14:169-187.

Franklin, S. 1997. Autonomous Agents as Embodied AI. *Cybernetics and Systems*. 28(6):499-520.

Franklin, S.; Kelemen, A.; and McCauley, L. 1998. IDA: A Cognitive Agent Architecture. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*. IEEE Press.

Freeman, W. J. 2003. A Neurobiological Theory of Meaning in Perception. Part I: Information and Meaning in Nonconvergent and Nonlocal Brain Dynamics. *International Journal of Bifurcation and Chaos*. 13(9):2493-2511.

Georgeon, O. L. and Ritter, F. E. 2012. An intrinsically-motivated schema mechanism to model and simulate emergent cognition. *Cognitive Systems Research*. 15:73-92.

Gerhart, J. and Kirschner, M. 2007. The Theory of Facilitated Variation. *PNAS*. 104(Supplement 1):8582-8589.

Gordon, I.; Vander Wyk, B. C.; Bennett, R. H.; Cordeaux, C.; Lucas, M. V.; Eilbott, J. A.; Zagoory-Sharon, O.; Jeckman, J. F.; Feldman, R.; and Pelphrey, K. A. 2013. Oxytocin enhances brain function in children with autism. *PNAS Early Edition*.

Greenspan, R. J. 2007. *An Introduction to Nervous Systems*. New York: Cold Spring Harbor Laboratory Press.

Hinchey, M. and Sterritt, R. 2012. 99% (Biological) Inspiration… In *Conquering Complexity*, M. Hinchey & L. Coyle eds., London: Springer-Verlage Ltd. Pages 177-190.

Kitano, H. 2002. Systems Biology: A Brief Overview. *Science*. 295:1662-1664.

Kuhn, T. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lighthill, J. 1986. The Recently Recognized Failure of Predictability in Newtonian Dynamics. *Proceedings of the Royal Society of London A: Mathematical, Physical, and Engineering Sciences*. 407(1832):35-50.

Lyon, P. 2006. The biogenic approach to cognition. *Cognitive Processing*. 7(1):11-29.

Marsh, K. L.; Johnston, L.; Richardson, M. J., and Schmidt, R. C. 2009. Toward a radically embodied, embedded social psychology. *European Journal of Social Psychology*. 39:1217-1225.

Marsh, K. L.; Richardson, M. J.; Baron, R. M.; and Schmidt, R. C. 2006. Contrasting Approaches to Perceiving and Acting With Others. *Ecological Psychology*. 18(1):1-38.

Maturana, H. R. and Varela, F. J. 1980. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht, Holland: D. Reidel Publishing Co.

Narasimharao, N.; Xi, L.; Bhattacharyya, S.; Widom, J.; Wang, J.-P.; Reeve, J. N.; Santangelo, T. J.; and Fondufe-Mittendorf, Y. N. 2013. Archael Nucleosome Positioning In Vivo and In Vitro is Directed by Primary Sequence Motifs. *BMC Genomics*. 14:391.

Peckham, H. E.; Thurman, R. E.; Fu, Y.; Stamatoyannopoulos, J. A.; Noble, W. S.; Struhl, K.; and Weng, Z. 2007. Nucleosome Positiong Signals in Genomic DNA. *Genome Research*. 17:1170-1177.

Peña, J.; Levy, R.; Hinchey, M.; and Ruiz-Cortés, A. 2012. Dealing with Complexity in Agent-Oriented Software Engineering: The Importance of Interactions. In *Conquering Complexity*, M. Hinchey & L. Coyle eds., London: Springer-Verlage Ltd. Pages 191-214.

Polanyi, M. (1968). Life's irreducible structure. *Science*. 160(3834):1308-1312.

Pollack, J. B. 1994. On Wings of Knowledge: A Review of Allen Newell's *Unified Theories of Cognition*. In *Contemplating Minds: A Forum for Artificial Intelligence*, ed. W. J. Clancey, S. W. Smoliar, and M. J. Stefik. Cambridge, MA: MIT Press. Pages 109-123.

Pyle, A. M. 2010. The Tertiary Structure of Group II Introns: Implications for Biological Function and Evolution. *Critical Reviews in Biochemistry and Molecular Biology*. 45(3):215-232.

Rosen, R. 1970. Structure and Functional Considerations in the Modelling of Biological Organization. Center for Theoretical Biology, State University of New York at Buffalo, 77(25): 1-12.

Rosen, R. 2000. *Essays on Life Itself*. New York: Columbia University Press.

Russell, P. and Norvig, S. J. 2003. *Artificial Intelligence: A Modern Approach*. 2nd edition. Upper Saddle River, NJ: Pearson Education, Inc.

Schrödinger, E. 1944. *What is Life?* Cambridge: Cambridge University Press.

Strain, S.; Franklin, S.; Heck, D.; and Baars, B. J. In preparation. Brain Rhythms, Cognitive Cycles, and Mental Moments: A New Approach in the Science of Cognition. In preparation.

Varela, F. ; Thompson, E.; and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: MIT Press.

Von Bertalanffy, L. 1950. The Theory of Open Systems in Physics and Biology. *Science*. 111:23-29.

Von Uexküll, J. 1926. *Theoretical Biology*. New York: Harcourt, Brace & Co.

Warnecke, T.; Batada, N. N.; and Hurst, L. D. 2008. The Impact of the Nucleosome Code on Protein-Coding Sequence Evolution in Yeast. *PLoS Genetics*. 4(11):e1000250.

Willingham, A. T. and Gingeras, T. R. 2006. TUF Love for "Junk" DNA. *Cell*. 125:1215-1220.