Journal of Artificial General Intelligence 3(2) 31-63, 2012
DOI: 10.2478/v10229-011-0016-2

Submitted 2012-06-07
Accepted 2012-10-07

Cognitive Architectures and Autonomy:

Commentary and Response

Editor: Włodzisław Duch, Ah-Hwee Tan, Stan Franklin

Autonomy for AGI

Cristiano Castelfranchi CRISTIANO.CASTELFRANCHI@ISTC.CNR.IT
ISTC-CNR
Italy

This paper provides a very useful and promising analysis and comparison of current architectures of autonomous intelligent systems acting in real time and specific contexts, with all their constraints. The chosen issue of Cognitive Architectures and Autonomy is really a challenge for AI current projects and future research.

I appreciate and endorse not only that challenge but many specific choices and claims; in particular: (i) that "autonomy" is a key concept for general intelligent systems; (ii) that "a core issue in cognitive architecture is the integration of cognitive processes ...."; (iii) the analysis of features and capabilities missing in current architectures; (iv) that an appropriate benchmark is still lacking; (v) the stress on "real time", on learning, on "resource management" (though goals and motivation would deserve a more central role); and especially (vi) the nice characterization of "some key features that a robot's mind must possess" (attention, expectation, models, ...).

1. What is "Autonomy"?

The main weakness I find in this work and perspective is the theory/analysis of "autonomy"; which would require a more analytical and systematic approach. What is autonomy? What are its components and dimensions? Which aspects of our cognitive and pragmatic architecture make us autonomous?

It is a pity in particular that the authors do not refer to the large debate on "Autonomy" in the '90s in the AI community of "Autonomous Agents", available either on the web ("Agent list") or in specific workshops and symposia (for ex. ATAL: Agents Theories Architectures and Languages; AAAI Symposium), or in several books and articles (See the References below on that debate). A deep and operational analysis of "autonomy" might have been very useful for this sort of AGI manifesto.

First of all, autonomy is a relational notion: X is autonomous and acts autonomously *from* something else; *from* some other "agent" (not necessarily a cognitive one).

There are two main kinds of autonomy:

- From the physical environment with its inputs and causal chains: X is not just automatically "responding" to stimuli. "Autonomous agent" means – in general – non hetero-directed; i.e. an agent whose behavior is not determined and driven from outside. Agents (at this level) are at least Goal-Oriented systems, not simply causal entities.

© 2012 Duch et al; Distributed under Creative Commons Attribution License

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

- From the social environment, that is from other cognitive-social agents able or willing to influence, control, or exploit the agent. With a very special and interesting issue for artificial agents: autonomy in "delegation", while acting "on behalf of" a given agent (for example the user, or the boss, etc.).

Second, autonomy has contents/dimensions; X may be autonomous (from Y) *"as for"* something but not for something else. X can be autonomous from Y as for a given ability/skill but not for another one; or autonomous as for perception or planning but not as for the freedom to act (permission, rights); or autonomous relative to a given resource but not another one; and so on. This provides a very articulated dimensional analysis of autonomy and of the social autonomy relations. This would also provide a specific picture of which components of cognitive (& motivational & action) architecture are crucial and make the agent autonomous.

A specific and very crucial dimension is "goal autonomy", implying that the system is self-interested, guided by its own internal motives, not just programmed from outside and "executing" orders (with only some "adaptation" freedom about the time of execution and the means to be used).

Another crucial dimension is "autonomy in believing": having one's own independent "sources" of information; being able to integrate knowledge and to form justified, supported beliefs (based on arguments and evidence). And on such a basis resisting possible wrong information or suggestions; grounding one's goal preferences and choice in one's autonomous beliefs.

2. Resources and Goals

The stress on "resource management" is very important for realistic systems, but unfortunately its characterization is weak and a bit biased, because the motivational aspects of the system (and of its autonomy) are somehow neglected. Resources are just "means" relative to, and subordinate to, goals. It is true that conflicts (and the need for choices and priorities) come from the needed resources, but the choice is between goals, and the priority is about goals (either instrumental or final ones).

References

AAAI Spring Symposium on Agents with Adjustable Autonomy, March 22-24, 1999, Stanford University.

Castelfranchi, C., 1995. Guaranties for Autonomy in Cognitive Agent Architecture. In M.J. Wooldridge and N. R. Jennings (eds.) *Intelligent Agents I*, Berlin, Springer.

Castelfranchi, C., 1999. What Autonomy Eventually is. Merging Agent Architecture, Theory of Action, and Dependence Theory for a Principled Analysis of Autonomy. Agents Theories Architectures and Languages (*ATAL 99*) Proceedings.

Castelfranchi C., Falcone R., 2003. From Automaticity to Autonomy: The Frontier of Artificial Agents. In: Hexmoor H, Castelfranchi, C., and Falcone R. (Eds), *Agent Autonomy*, Kluwer Publisher, pp.103-136.

Falcone R. and Castelfranchi C., 1997. "On behalf of ..": levels of help, levels of delegation and their conflicts, *4th ModelAge Workshop: "Formal Model of Agents"*, Siena, Italy.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

Franklin S. and Graesser A., 1997. Is it an Agent, or just a Program: A Taxonomy for Autonomous Agents. In J.P. Muller, M.J. Wooldridge, N.R. Jennings (eds.) *Intelligent Agents III*, Berlin, Springer, LNAI 1193.

Luck, M. and d'Inverno M., 1995. A Formal Framework for Agency and Autonomy. *ICMAS-95*, pp. 254-60

Maes P., 1990. Situated agents can have goals. In P. Maes, editor, *Designing Autonomous Agents*, pp. 49-70. The MIT Press.

Wooldridge M., and Jennings N., 1995. Intelligent Agents: Theory and Practice, *The Knowledge Engineering Review*, Vol. 10, N.2, pp. 115-152.

Are Disembodied Agents Really Autonomous?

Antonio Chella ANTONIO. CHELLA@UNIPA.IT
DICGIM - University of Palermo
Viale delle Scienze, building 6
90128 Palermo, Italy

The target paper by Thórisson and Helgasson is an informed review of some cognitive architectures and it also proposes valuable comparisons among the reviewed architectures. However, the main problem unsolved by the paper is: when an agent could be really considered autonomous?

According to the target paper: "Autonomous systems automatically perform tasks in some environment, with unforeseen variations occurring in both, through some type of automatic learning and adaptation that improves the system's performance with respect to its high-level goals." (Thórisson and Helgasson, pag. 3).

Frankly, this definition seems to be a sort of smart shortcut: instead of describing the overall behaviors of a good old-fashioned control architecture in terms of algorithms, processes and message passing protocols, the designers of an autonomous architecture typically inject heuristic thumb-rules disguised as "cognitive" modules into the agents. The results are sometimes amazing but this is not necessarily a real progress in understanding what autonomy is, and how we can manage it: it is only wishful labeling of boxes.

We claim that a basic challenge for a system to be seriously autonomous is that the agent should be embodied in a broad sense: i.e., it should have a body and it should be able to employ its own body to sustain long-time tight interactions with the external environment to pursue its own goals (see Chella and Manzotti, 2009).

Noteworthy, the general challenge of embodiment is far more complex than the simple idea of moving the robot while controlling actuators and sensors, as reported in Fig. 1 of the target paper. Embodiment refers to the kind of development and causal processes engaged between the agent, its body, and its external environment.

A real autonomous agent is in fact constrained to continuously interact with the environment: it acts in the external world (which could be completely unknown, as the Martian world discussed

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

in the target paper) and it receives continuous feedbacks from the world. The world itself provides the ground for the internal symbols of the agent.

Therefore, the design of a real autonomous agent cannot underestimate the importance of the body. To avoid embodiment in the design of an autonomous architecture may mean to not capture the main concept of autonomy.

Yet we admit that embodiment is a deep and complex challenge. On one hand, there is no such thing as a "non embodied agent", since even the most classic AI system is implemented as a physical set of instructions running inside a physical device. On the other hand, even the more complex state-of-the-art robots, such as ASIMO[1] or the Geminoid[2] are controlled by a suitable set of classic controller loops whose behavioral rules are hard-wired by designers.

Some biological agents would apparently score very well on embodiment, but they do not seem good candidates for being called autonomous agents: take insects, for instance. They show impressive morphological structures that allow them to perform outstandingly well, but obviously they do not have any cognitive capabilities.

On the other side, having a body may also influence higher cognitive processes. In their seminal and debated book, Lakoff and Núñez (2000) discuss in detail and with many examples how mathematical concepts and reasoning are rooted in the human body and in its interactions with its environment.

The target paper pointed out that an agent does not really need to be embodied in the real world: it could also operate in a virtual world, like a Second Life[3] avatar. However, we do not have any empirical evidence that such a "non embodied situated brain" would ever be really autonomous.

What kind of architecture is sufficient for an embodied situated agent? The animal literature presents interesting cases regarded as examples of tight integration with the environment since they outsource part of their control processes by smart bodily arrangements, as for example in the case of the control circuitry of the octopus arm, which is embedded in the arm itself (Sumbre et al., 2001; see also Chapter 3 of Franklin, 1995). In the robotics literature, Paul (2006) introduced the notion of "morphological computation", in order to stress the tight relationships between the morphology and the control requirements of a robot.

Following this line of thinking, we claim, in agreement with Franklin and Graesser (1997), that the real challenge of autonomous agency involves the concept of developmental integration with the environment such that what the robot is and does is a result of the present and past interactions with the environment. We therefore propose an alternate definition of autonomous agent as a strongly non Markovian agent able to change its internal structure, and possibly its external shape (as in the case of self-reconfigurable robots, see Yim et al. 2007) in non-trivial ways as a result of its tight coupling with the environment during its own operating life. Also the agent's own autonomy may evolve during this whole-life integration.

How embodiment in this broad sense may be employed for an autonomous robot acting as an explorer of Mars? Two examples briefly mentioned below are related to resilient and energetic autonomy.

An example of resilient autonomy is the starfish robot by Bongard et al. (2006). This robot is able to build from scratch an inner model of its own body by moving itself in a random way and by registering the responses of movements from the environment by suitable sensors. The starfish body model is resilient in the sense that it updates its own body model consequently if some

1 http://asimo.honda.com/
2 http://www.geminoid.jp/
3 http://secondlife.com/

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

damages occur to the robot's body. This kind of resilient autonomy is perfectly suitable for the Martian exploratory task proposed in the target paper.

An example of energetic autonomy is Ecobot, discussed by Melhuish et al. (2006). Intriguingly, Ecobot is an energetically autonomous robot able to search for suitable insect biomass in the environment and to convert the biomass into energy to refuel the robot itself. This is a real example of autonomy in a physical sense: the robot needs to think and act to charge its own batteries in a non trivial way in order to continue its operations. Also in this case, this kind of autonomy is suitable for the discussed exploratory task.

Instead, the architectures reviewed in the target paper lack these kinds of real autonomy: they remain more or less the same notwithstanding their interplay with their environment. Their internal and external structure are largely unchanged by their interactions with the environment.

It should be noticed one more time that the issues related with embodiment are deep scientific and theoretical problems and not simply technical challenges. The emerging general picture is that of the needs of a novel, yet-to-be-defined framework for embodied autonomous agents.

A final consideration: the target paper seems to be largely rooted in the general mainstream of cognitive sciences. Now, roughly, the background hypothesis of cognitive science is that the Turing machine is, with some distinctions, a valid model of the mind. In the centenary celebration year of Alan Turing, one may wonder if it is now time to take into account some different model of cognition that takes into account the body, the mind and the external environment.

References

Chella, A.; and Manzotti, R. 2009. Machine Consciousness: A Manifesto for Robotics. *International Journal of Machine Consciousness*. 1: 33 – 51.

Lakoff, G.; and Núñez, R.E. 2000. *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being.* New York: Basic Books.

Sumbre, G.; Gutfreund, Y.; Fiorito, G.; Flash, T.; and Hochner, B. 2001. Control of Octopus Arm Extension by a Peripheral Motor Program. *Science*. 293: 1845 – 1848.

Franklin, S. 1995. *Artificial Minds.* Cambridge, MA: MIT Press, Bradford Books.

Paul, C. 2006. Morphological Computation. *Robotics and Autonomous Systems*. 54: 619 – 630.

Franklin, S.; and Graesser, S. 1997. Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent Agents III Agent Theories, Architectures, and Languages* ed. J. Müller and N. Jennings. Lecture Notes in Computer Science, 1193: 21 – 35. Berlin, Heidelberg: Springer.

Yim, M.; Shen, W.-M.; Salemi, B.; Rus, D.; Moll, M.; Lipson, H.; Klavins, E.; and Chirikjian, G. S. 2007. Modular Self-Reconfigurable Robot Systems. *IEEE Robotics & Automation Magazine*. March: 43 – 52.

Bongard, J.; Zykov, V.; and Lipson, H. 2006. Resilient Machines Through Continuous Self-Modeling. *Science.* 314: 1118 – 1121.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

Melhuish, C.; Ieropoulos, I.; Greenman, J.; and Horsfield, I. 2006. Energetically autonomous robots: Food for thought. *Autonomous Robots*. 21: 187 – 198.

The perception-…-action cycle cognitive architecture and autonomy: the view from the brain

Vassilis Cutsuridis VCUTSURIDIS@GMAIL.COM
Division of Engineering
Kings College London
Strand, WC2R 2LS
U.K.

Autonomous systems with reasoning capabilities are systems able to perceive their environment and act on it by performing complex tasks automatically. Autonomous systems are also able to adapt to unforeseen operating conditions or errors in a robust and predictable manner without the need for human guidance, instructions or programming. To accomplish such complex feats they must master the powers of perception, recognition, attention, learning and memory, cognitive control, reward and motivation, decision making, affordance extraction, action planning and action execution (step 1). Once these powers are successfully mastered, then these systems may be embodied into a robot able to act in the real world (step 2). Their embodiment, however, cannot guarantee that these systems will be able to operate autonomously in the environment as they will still need to solve the issues of the real-time system operation, resource management and meta-learning (step 2).

In their article "Cognitive architectures and autonomy: a comparative review" Thórisson and Helgasson reviewed a number of "autonomous" systems and architectures with general "cognitive" capabilities and compared and contrasted their performance in a hypothetical example of autonomous exploration of an environment by a robot. Instead of their criteria focusing on how the powers of perception, recognition, attention, memory, cognition, decision making and action planning and execution are achieved by these systems (step 1), the authors ignored these powers, and compared and contrasted the systems based on step 2's real-time processing, resource management, learning and meta-learning issues. The authors argued that the former functions (e.g. perception, recognition, attention, memory, etc.) are less important.

I believe that dealing with the issues of real-time processing, resource management, learning and meta-learning first and comparing and contrasting the reasoning capabilities of systems based on them is similar to building a house from the roof down. The systems are forced to solve the real-time system operations of functionalities which they have not deciphered yet, so they will inevitably be dumb, as they will be empty shells not possessing any reasoning powers that will enable them to go beyond the information provided.

Furthermore, though some of the reviewed systems are "biologically inspired" in that they depend on behavioral studies and test themselves by the replication of experimental behavioral data, none of these systems attempt reverse engineering of the brain circuitry that supports these behaviors. Reasoning is the highest faculty of the human brain and it depends on the majority of the brain components (perception, attention, learning and memory, decision making, action, etc.). The brain is a system that has evolved over a million or so years, so it is expected to provide a reasonably optimized solution to many of the cognitive tasks under consideration.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

I propose as an alternative to the systems reviewed by the authors a brain-inspired cognitive control architecture for autonomous interaction of a robot with objects situated in its immediate environment (i.e. a form of exploration of an environment by a robot). My approach is based on work done in the EU-DARWIN project. A graphical representation of the cognitive control architecture is given in Figure 1 (Cutsuridis, 2012). The architecture proposes that exploring an environment requires to act upon objects in it, like in the case of vision-guided reaching and grasping of objects. The objects themselves are not to be known a-priori to the system, but their knowledge is built by the system through interaction and experimentation with them. The architecture is multi-modular, consisting of object recognition, object localization, attention, cognitive control, affordance extraction, value, decision making, motor planning and execution modules. The components of the architecture are novel as well as based on previous architectures (Cutsuridis et al., 2011; Cutsuridis, 2009; Taylor et al., 2009) and follow very closely what is currently known of the human and animal brain.

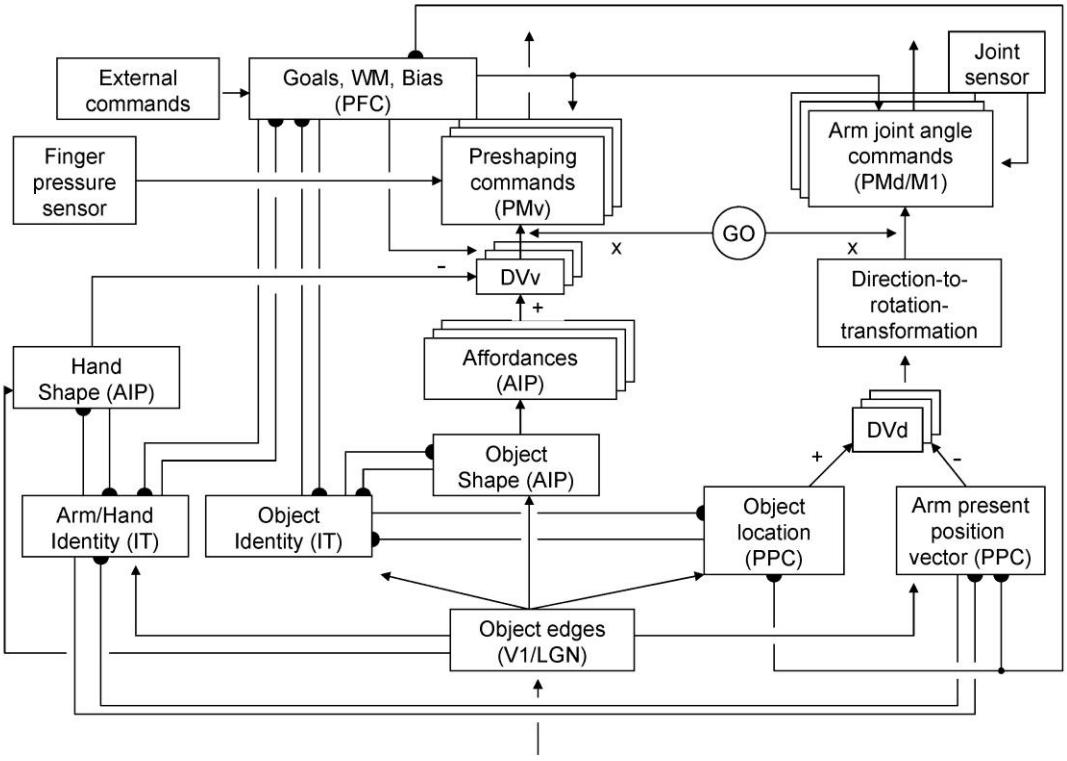

Figure 1. Graphical representation of the cognitive control architecture of object shape and object location recognition, attention reward, decision making, cognitive control, affordances, action planning and execution in reaching and grasping.

Vision-guided reaching and grasping involves two separate visuomotor channels, one for reaching and another one for grasping, which are activated in parallel by specific visual inputs and each channel controls specific parts of limb (arm and hand, respectively). An input image is processed in a bottom-up fashion, providing input to feature detectors, which in turn lead to the formation of visual maps (the *object* map and the *spatial saliency* map). Bidirectional cross-talk

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

between object and spatial maps ensures that the object corresponds to the appropriate spatial location in the environment. The visual maps then activate the *cognitive control* map (goals, motivations, task constraints), which in turn feeds back to amplify the neural representations in the visual maps, which are relevant to the current context, and to suppress the irrelevant ones. Resonance between goals and object, and goals and spatial maps is achieved via a measure of degree of similarity, which depends on the amount of modulation (value) the maps receive from the dopamine system. A winner-take-all competition between resonated neural representations ensures that the object representation and spatial representation that reached resonance first will continue fastest processing first, followed by the second fastest and so on. Once an object and a spatial representation is selected a library of action plans are selected, one for reaching and the other one for grasping. Once again the cognitive control maps will select the action plan most relevant to the current context and suppress the irrelevant ones. The selected reaching and grasping motor plans will be gated by a GO signal (output of the basal ganglia), and form the final motor commands, which will be sent to the motor execution centers for execution. Visual and proprioceptive feedback will update the current arm position and fingers configuration towards the desired ones.

My cognitive control system has been implemented on the iCub robot with considerable success when multiple objects were situated in the environment and the robot had to recognize them, localize them, attend to each of them and reach and grasp them according to an externally dictated sequence of motor actions.

References

Cutsuridis, V.; Hussain, A.; and Taylor, J.G. 2011. *Perception-action cycle: Models, architectures and hardware*. New York, NY: Springer

Cutsuridis, V. 2009. A cognitive model of saliency, attention and active visual search. *Cognitive Computation*. 4(1): 292-299

Cutsuridis, V. 2012. A cognitive control architecture of the perception-action cycle for robots and agents. *Under review*.

Taylor, J.G.; Hartley, M.; Taylor, N.R.; Panchev, C.; and Kasderidis, S. 2009. A hierarchical attention-based neural network architecture, based on human brain guidance, for perception, conceptualisation, action and reasoning. *Image and Vision Computing* 27(11):1641–1657

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

Autonomy Requires Creativity and Meta-Learning

Włodzisław Duch WDUCH@IS.UMK.PL
Department of Informatics,
Nicolaus Copernicus University,
Toruń, Poland,
&
School of Computer Engineering,
Nanyang Technological University,
Singapore

1. Introduction

Recently several review articles on cognitive architectures (CAs) have been published by Hui-Qing, Tan, and Ng (2007); Langley, Laird, and Rogers (2008); Duch, Oentaryo, and Pasquier (2008); Taatgen and Anderson (2009), and de Garis et al. (2010). All these articles, as well as the current target article, ignore some important issues related to creativity, imagination, intuition and insight. The target article is focused on relations between cognitive architectures and autonomy, stressing four main aspects: real-time system operation, resource management, learning, and meta-learning. These aspects are certainly important, although one may argue that autonomous internet agents – for example spiders that explore scientific journal sites, collecting relevant research results and using this information to create models of biological organisms (see the *biocyc.org* for some examples) – should rather be based on cognitive architectures that are of a different kind. Such agents do not have to operate in real time, do not require full-blown embodied CAs that may be necessary for robotics, they may use only a subset of all functions that robots need. However, they are goal-directed, have to plan their actions, perceive relevant information in many forms, understand it in relation to already accumulated knowledge, and work in an autonomous fashion. Their domain of competence is detached from concepts that acquire meaning from direct embodiment, although some would argue that these agents are embodied (Franklin 1997). Such architectures are also of great interest to the AI community.

2. Meta Learning

Meta-learning is defined in the target article as "the ability of learning to learn". In agreement with the recent approaches to meta-learning in computational intelligence, presented in the Jankowski, Duch, and Grąbczewski (2011) book, non-linear learning rates advocated in the target paper are certainly not sufficient to reconfigure the whole system in a flexible way. Learning may help to improve tasks for which the systems is designed, but it is individual development that helps to structure the whole design. The evo-devo, or evolutionary core of skills combined with the developmental perspective, captures these two steps rather well. The third step is neural reuse, flexibility in reconfiguring brain modules at different levels to solve complex tasks, as Anderson (2010) has recently discussed. This type of learning systems, searching for interesting solutions in the space of all possible models, are quite complex, but have a chance to become a new generation of learning tools. In a way they use their introspective-like capabilities to combine the

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

existing modules that are already doing useful work in a novel way, transferring knowledge between different tasks, creating new emergent higher-level modules that may reconfigure internal information flow enabling new functions. In this respect, as the authors have pointed out, existing cognitive architectures have not made much progress. However, Artificial General Intelligence requires meta-learning, and this aspect, connected to creativity and autonomy, has not yet been sufficiently appreciated.

Solving novel, unforeseen problems, is a key ingredient of intelligent autonomous systems. A single ant is partially autonomous, exploring the environment in a blind way and cooperating with other ants, but it has a rather rigid behavior. Blind Variation Selective Retention (BVSR) principle has been the basis of psychological theories of creativity for over 50 years, as reviewed by Simonton (2010). Blind variation is not random, it is biased in many ways by the tools (including the body) and past experiences (reflected in probability of different brain activations leading to various associations), while selective retention is biased by values, needs, goals and preferences. The space of all imaginings is constrained by patterns that may potentially arise in systems that have some knowledge, encoded either in symbolic form, or incorporated in the subsymbolic neural patterns. In networks imagination arises from distributed fluctuating neural activity, constrained by the strength of associations between subnetworks coding different concepts. These patterns compete with each other and those that are most interesting – lead to emotional arousal, have most associations, provide steps towards the final goal – are retained, amplifying certain associations and discovering partial solutions that may be useful in view of the set goals. This process is biased by many context layers (species, culture and language-specific) priming expectations. Results of various sequences of actions may be imagined before the action is taken and their utility estimated.

3. AI and Creativity

AI systems approximate this process using heuristic search procedures. Is the symbolic approximation of what may be encoded in the network activations sufficient? Intuitive behavior based on experience that cannot be easily verbalized with a finite number of rules may be captured in various types of networks. The Numenta Grok prediction engine, based on hierarchical temporal memory cortical learning algorithm of Hawkins (2011) works in real time finding spatial and temporal patterns in streams of data and anticipating consequences. Although it is a long way from becoming a full brain-inspired cognitive architecture, it is an interesting step towards a better approximation of brain processes for real-time control of autonomous systems.

Creativity permeates our everyday actions. Autonomous systems continuously solve problems that require novel sequences of movements, actions and thoughts. Understanding language, including artificial programming languages, is a good example here, understanding people's intentions and interactions is another. Although very few people have sufficient scientific knowledge to create new theories that are worth Nobel prizes, all humans (and many animals) are creative in the sense of understanding (at least most of the time) and manifesting novel expressions or behaviors. Can autonomy be achieved without creativity? Brain mechanisms behind creativity, imagination, intuition and insights are a consequence of information processing in complex brain networks and can be captured in computational models, as described by Duch (2008). However, so far cognitive architectures have not yet incorporated them, although they may be important on the road to full autonomy.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

References

Anderson, M. 2010. Neural reuse: A fundamental organizational principle of the brain. Behavioral and Brain Sciences 33(4):245–313.

de Garis, H.; Chen, C.; Goertzel, B.; and Lian, R. 2010. A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures. Neurocomputing 1-3:30–49.

Duch W.; Oentaryo R. J.; and Pasquier M. 2008. *Cognitive architectures: where do we go from here?* In: Frontiers in Artificial Intelligence and Applications, Vol. 171 (Ed. by Pei Wang, Ben Goertzel, and Stan Franklin), IOS Press, pp. 122-136.

Duch, W. 2008. Intuition, Insight, Imagination and Creativity. IEEE Computational Intelligence Magazine 2(3):40–52.

Franklin, S. 1997. Autonomous Agents as Embodied AI. Cybernetics and Systems 28:499–520.

Hawkins, J. 2011. Hierarchical Temporal Memory including Cortical Learning Algorithms. Technical report, Numenta, Inc. Ver. 0.2.1.

Chong, H.-Q.; Tan, A.-H.; and Ng, G.-W. 2007. Integrated cognitive architectures: a survey. Artificial Intelligence Reviews 28:103–130.

Jankowski, N.; Duch, W.; and Grąbczewski, K., eds. 2011. Meta-learning in Computational Intelligence., volume 358 of Studies in Computational Intelligence. Springer.

Langley, P.; Laird, J.; and Rogers, S. 2008. Cognitive architectures: Research issues and challenges. Cognitive Systems Research 10:141–160.

Simonton, D. 2010. Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. Physics of Life Reviews 7:156–179.

Taatgen, N.; and Anderson, J. 2009. The Past, Present, and Future of Cognitive Architectures. Topics in Cognitive Science 1:1–12.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

Meta Learning, Change of Internal Workings, and LIDA

Ryan McCall RMCCALL@MEMPHIS.EDU
Fedex Institute of Technology 403h
The University of Memphis
Memphis, TN 38152, USA

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

1. Introduction

The authors review a number of well-known cognitive architectures, as well as several explicitly chosen for their integrated nature and broad scope. Recently there have been other comparisons of cognitive architectures including one by Duch et al. (2008) in addition to the BICA table[1].

2. Meta Learning and Change of Internal Workings

A novel aspect of this review is its focus on autonomy. The authors identify "meta learning" as a necessary sub-aspect of autonomy, and use it as a metric for comparing architectures. There are several issues with this process. First, no definition or defining references in the initial explanation of meta learning are given. However, this term is used in multiple contexts, e.g. metacognition, where it is called metacognitive regulation[2], and also in the context of machine learning (e.g. Schaul & Schmidhuber 2010). More clarification on the nature of "meta learning" is needed. What aspects of "meta learning" can be accomplished by biological agents (e.g. humans)? What aspects of "meta learning" go beyond the capabilities of biological agents?

While there is ample evidence that the ability to monitor and regulate one's learning is absolutely critical – at least in humans, it seems stronger to suggest that an agent might "change its own internal workings." Does this include major changes to the agent's architecture? There doesn't seem to be any justification for such drastic self-modification being necessary or even valuable. Has there been a demonstration that an agent changing its "own internal workings" in real time has improved performance? It seems that if an agent changes its own architecture online, there could be a catastrophic failure resulting in the agent becoming inoperable. For example a change in "internal workings" of a human brain could produce epilepsy rendering that human dysfunctional.

Why is meta-learning, as defined, a research priority when a working human-level AGI system has not yet been produced? It could be argued that the best meta learning approach for the foreseeable future is for humans to continue researching AGI architectures, and generating and testing various designs. No AGI architecture seems remotely close to be competing with humans on AGI design capabilities at this time.

1 http://bicasociety.org/cogarch/architectures.htm
2 http://en.wikipedia.org/wiki/Metacognition

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

3. Clarifying Issues with the LIDA Architecture

This portion of this comment is dedicated to explicating misimpressions of the LIDA model.

Describing LIDA sec. 3.8 of the target paper reads "... availability of resources does not affect the processing of the system, with each operating cycle always selecting a single coalition of data for further processing." Though the second clause of the quoted sentence is correct, the first clause does not follow from it. Here are four examples: 1) Even over a single cognitive cycle, availability of resources would typically affect the determination of saliency (presence, importance, urgency, insistence, novelty, unexpectedness, loudness, brightness, motion, etc.), the currency of the competition for consciousness. For example a situation recognized as urgent (perhaps due to a limited time situation) will have increased saliency thus increasing its chance to be attended upon by the consciousness mechanism. 2) Over multiple cycles, LIDA has long implemented, at least conceptually, deliberation. Given a circumstance with multiple appropriate options (information overload) LIDA performs volitional decision making via James' ideomotor theory (1890). LIDA's implementation of volitional decision making employs a timekeeper whose time window determines when the volition must stop, and a decision be made (Franklin 2000). 3) Since schemes are chosen for instantiations largely by the intersection of their context with the current conscious broadcast, the availability of resources plays a critical role in deciding which schemes in Procedural Memory are instantiated into behaviors to be considered by the Action Selection module. 4) Finally, recent work on LIDA modeling the attentional blink phenomenon (Madl, Franklin, 2012) has employed a finite attention resource that may be temporarily "used up" by demanding percepts.

On the same page we find "...the many different types of learning supported by the architecture, both symbolic (e.g. declarative) and sub-symbolic (e.g. perceptual)." The LIDA model is in no sense symbolic. Rather it would come under the category of embodied (situated) cognition, as it implements Barsalou's perceptual symbol system in its Perceptual Associative Memory. One might classify LIDA as subsymbolic in that activation passing occurs throughout. However, there are no artificial neural networks, as such, implemented in LIDA. So, the term "sub-symbolic" might also be a little misleading.

There are also issues with the assessment of LIDA along the "Real-time" dimension. Ticks are not "operating cycles", that is, they are not directly tied to the cognitive cycle. Ticks are a discretization of time for purpose of computer implementation of LIDA's processes, and are specific to the LIDA computational framework. One tick can take an arbitrarily short amount of real time given a sufficiently powerful computer. Ticks are not a part of the conceptual LIDA model, however the processes of the LIDA model do differ in their real time length, which has consequences for LIDA's real-time performance. To further illustrate: to achieve a cognitive cycle of 300 ticks there may be several processes running at various frequencies e.g. 20, 100, 50 etc. Please see Snaider, McCall, & Franklin (2011) for an extended discussion of ticks and the larger LIDA computational framework. Madl, Baars, & Franklin (2011) gives an example of an agent implementation using ticks.

Central to the LIDA model is the cognitive cycle. LIDA's cognitive cycles cascade, that is, they overlap. As a model of human cognition LIDA predicts that three different such sense-action cycles could be occurring concurrently (each in a different processing stage) (Madl, Baars, Franklin, 2011). The only bottleneck in the cognitive cycle comes from the consciousness mechanism. Here there is a minimum amount of time that must pass (~100ms in humans) between successive conscious broadcasts which functions to provide stability and to preserve seriality.

Time can be involved in behavior streams implementing higher order cognitive processes. For example, as mentioned earlier, LIDA performs volitional decision making via James' ideomotor

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

theory (1890), using a timekeeper whose time window determines when volition must stop and a decision be made (Franklin 2000). This is one example of time as a resource that may affect LIDA's cognitive processing. Another example is in LIDA's consciousness mechanism where a time-based trigger may initiate a competition for consciousness if a certain amount of time has passed with no conscious broadcast occurring (see Madl, Baars and Franklin 2011).

References

Duch W.; Oentaryo R. J.; and Pasquier M. 2008. *Cognitive architectures: where do we go from here?* In: Frontiers in Artificial Intelligence and Applications, Vol. 171 (Ed. by Pei Wang, Ben Goertzel, and Stan Franklin), IOS Press, pp. 122-136.

Franklin, S. 2000. Deliberation and Voluntary Action in 'Conscious' Software Agents. *Neural Network World, 10*, 505–521

James, W. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Madl, T.; Baars, B. J.; and Franklin, S. 2011. The Timing of the Cognitive Cycle. *PLoS ONE, 6*(4), e14803.

Madl, T.; and Franklin, S. 2012. *A LIDA-based Model of the Attentional Blink.* Paper presented at the 11th International Conference on Cognitive Modeling, Berlin, April 13-15.

Schaul, T.; and Schmidhuber, J. 2010. Metalearning. *Scholarpedia, 5*(6):4650.

Snaider, J.; McCall, R.; and Franklin, S. 2010. *The Immediate Present Train Model Time Production and Representation for Cognitive Agents*. Paper presented at the AAAI Spring Symposium on "It's All In the Timing", Palo Alto, CA.

Snaider, J.; McCall, R.; and Franklin, S. 2011. *The LIDA Framework as a General Tool for AGI*. The Proceedings of the Fourth Conference on Artificial General Intelligence (AGI-11).

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

An Appeal for Declaring Research Goals

Brandon Rohrer BROHRER@GMAIL.COM
Sandia National Laboratories
Albuquerque, NM 87185, USA

In their review, the authors present a challenging task, exploration of the natural universe, and assert that performing it would fulfill the main goals of AGI. From this challenge, they distill four sub-goals that they consider necessary to achieve it – real time, learning, resource management, and meta-learning – and use these as criteria for comparing several cognitive architectures.[1] The resulting autonomy-centered comparison provides valuable insights into the relative capabilities of different types of architectures. However, what they do not mention is that autonomy is not the only motivator for creators of cognitive architectures. The explicit goal of a number of architectures is to model the operation of the human brain.

In an informal show-of-hands survey at the 2009 Biologically Inspired Cognitive Architectures Symposium, the audience (consisting primarily of cognitive architecture creators) was asked to choose one of two hypothetical outcomes for their research: 1) Create a system that equaled human performance in all areas, but provided no insights into the operation of the brain, and 2) Create a system that demonstrably modeled every aspect of the brain, but could do nothing. Approximately half the respondents chose each option. Informal communications with a number of researchers since then have supported these results. Within the AGI community, there are strong emphases on both modeling and autonomy.

While this was a patently non-rigorous survey, not to mention a false dichotomy, it does illustrate the range of motivators in creating cognitive architectures. In fact, detailed conversations with practitioners support the notion that there are as many motivators for creating AGIs (and thus as many definitions of "success") as there are researchers. Comparing architectures on any one measure is instructive, but partial. This situation can be frustrating. With such orthogonal goals, modelers and autonomists often have difficulty communicating. And without explicit goals, presentations of model-focused architectures resemble beauty contests, and demonstrations of autonomy-focused architectures play out like a diverting series of pet tricks.

One way to address this is to be painfully explicit about our own goals. If we can definitively answer the questions **"What are you trying to do?"** and **"How do you know if you're getting better at it?"** we give clarity to our work. The answers are still useful, even if they change over time and from researcher to researcher. They provide a measure of progress by allowing us to compare our work to each other and to our past selves.

The more specific we can be in our goals, the more they will benefit us. For modeling-focused architectures, What is the scope of your modeling? Which phenomena are you trying to model? Which data sets will you compare your performance to? For autonomy-focused architectures, what task space are you planning to operate in? Which tasks are you trying to

1 Disclaimer: I am also the author of a cognitive architecture, BECCA. Like the authors of the target article, my own goals are strongly rooted in autonomy, focused on natural world interaction, (Rohrer, 2010) that is, creating a system that can do everything I can do. The intermediate tasks that I'm working toward are detailed in (Rohrer, 2010), the benchmark I am currently testing BECCA against is described in (Rohrer, 2012), and its Python code is downloadable from (Chapman, 2012).

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

perform? How will you measure your performance on them? If we can represent our architectures' fitness with a number, we can easily demonstrate their improvement over time.

It is true that clarity exposes us to some risk. After all, what if we fail to make progress? Or what if someone outperforms us on our own tasks? The benefits of clarity are too great to ignore. Clear goals keep our research focused, help us to allocate scarce resources to maximum effect, and simplify our communication. Being able to directly compare cognitive architectures provides a coherence to the field that increases its perceived rigor and legitimacy. And perhaps most importantly, specific goals help us to concisely describe our vision, helping us to capture the imagination of a new generation of researchers.

Acknowledgements

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

Chapman, M. 2012. BECCA home page. Available electronically at http://www.openbecca.org.

Rohrer, B. 2010. Accelerating Progress in Artificial General Intelligence: Choosing a Benchmark for Natural World Interaction. *Journal of Artificial General Intelligence* 2:1–28.

Rohrer, B. 2012. BECCA 0.4.0 User's Guide. Available electronically at http://www.openbecca.org.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

The Development of Cognition as the Basis for Autonomy

Frank van der Velde F.VANDERVELDE@UTWENTE.NL
Technical Cognition,
Centre for Telematics and Information Technology, Cognitive Psychology and Ergonomics,
University of Twente
PObox 217, 7500 AE Enschede
The Netherlands
IOP, Leiden University
The Netherlands

1. Introduction

In their paper, Thórisson and Helgasson (henceforth: TH) discuss the importance of autonomy for general artificial intelligence. They also analyze a number of cognitive architectures on four dimensions of autonomy, with most of them failing on at least one of them. The issues discussed by TH are important and the points they raise make a lot of sense. I find myself agreeing with their discussion on the need for autonomy, the importance of the dimensions of real time operation, learning, resource management and meta-learning, and their analysis of the cognitive architectures.

However, there is also something missing. TH's analysis of autonomy sounds rather cognitively naive. To illustrate this, consider TH's description of autonomy (page 4): "Let us imagine an exploration robot that can be deployed, without special preparation, into virtually any environment, and move between them without serious problems". Among the environments listed are Mars, the Sahara desert, the Amazon jungle and the depths of the ocean.

What would it take for a human to achieve this aim? First of all, no human would go to Mars without special preparation. Astronauts get extensive training and for good reason. Also, a city slicker would hardly be able to survive on his own in the Amazon forest or the Sahara desert. So, 'without special preparation' needs to be taken with a grain of salt.

But, more importantly, a human would need some 20 years or so of development and learning before he or she could even begin to explore these environments. Development of this kind is not covered by the learning dimension in the way described by TH. Development, certainly in early life, is much more than making a selection between different types of reinforcement learning or logical inference. The difference between these forms of learning and development can already be seen by noting that the other dimensions of autonomy (real time, resource management and meta-learning) are quite limited when cognitive development is at its peak.

2. Development and learning

A major difference between cognitive development and learning as described by TH concerns the effect of development on the structure of the cognitive architecture. In humans there seems to be

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

a close relation between cognitive development and brain development. This may be a reason for why it takes so much time to reach the level of cognitive adulthood.

A case in point is the development of language. On the one hand, the structure of the brain needs to reach a certain level of complexity before language can be learned (a stage never reached by the non-human primate brain). But on the other hand, language learning itself influences brain structure. If language is not learned at a certain age (around 12) it cannot be learned anymore at the level of full-blown natural language (e.g., Calvin and Bickerton, 2000). So, the development of language and the development of the brain go hand in hand. The brain determines language learning but language learning itself influences the structure of the brain.

The relation between brain development and language learning is highly complex, as exemplified by the observation that natural language cannot be learned by non-human primates (despite several attempt to do so). So, apparently, we need some basic brain structure to begin with and the ability for structural change and growth. But we also need the interaction with the environment and other language users for the language architecture to develop. That is, we need linguistic experience as well.

The interaction between (initial) brain structure, plasticity and linguistic experience, and the constraints they put on each other, determine the development of the architecture for language. This may be the reason why this development seems to proceed in stages (e.g., Saxton, 2010). You need a basic set of words before you can make basic (two or three word) sentences with them. In turn, you need these basic sentences before you can create and understand more complex sentences. There is a distinct possibility that these stages are accompanied by structural changes in the underlying architecture down to the 'hardware' level. In the case of human development one can indeed assume that cognitive development shapes the connection structure of the brain in a step by step manner, in which each step determines the potential for development of the next step. The existence of a critical period in which language learning has to occur underlines this point. When the brain is highly plastic, its connection structure can be molded by experience to develop into a language architecture. But when the brain's plasticity declines and a language architecture has not yet developed, the resulting brain structure cannot develop into a language architecture anymore.

Notice that this is more than just learning. In fact, our ability to learn may depend on this kind of development. We can learn throughout life, but the high plasticity underlying brain development in early life seems to reduce at the beginning of adulthood (as exemplified by the end of the critical periods for learning language of for the development of visual perception). There may be a sound reason for this: high plasticity is good when you need to develop a cognitive architecture. But once the architecture has developed, high plasticity can have adverse effects. It might result in a (too substantial) loss of acquired knowledge and abilities, when they are washed away by new experience.

3. Cognitive autonomy and development

So, what do we need to reach the lofty goal set out by TH? I would argue that, next to the dimensions discussed by TH, we also need to understand how a cognitive architecture can develop by interacting with its environment. This could entail a structural development that would shape the architecture stage by stage, in which each stage is needed for the development of the next one. This structural development of the cognitive architecture could even proceed down to the hardware level. TH does not really discuss hardware issues, but there are sound reasons to

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

assume that they will be important for arriving at truly autonomous generally intelligent systems. The increase of processing speed of single-core processors has come to an end. We need parallel architectures, with more than a few hundred processing cores. Likely, these processors and their interactions will require new forms of hardware (graphical processing units may be a step in this direction). Currently it takes a supercomputer to level human performance in chess or jeopardy, but it is difficult to see how such systems could be used to explore environments like Mars.

With new forms of parallel hardware, new forms of cognitive representations will likely arise, more resembling the neural assembly representations formed in the brain (e.g., Harris, 2005). In turn, these neural assembly representations, developed through experience (e.g. Hebb, 1949), will likely require different kinds of architectures for the development of linguistic and cognitive capabilities (e.g., van der Velde and de Kamps, 2006; 2010).

References

Calvin, W. H.; and Bickerton, D. 2000. *Lingua ex Machina*. Cambridge, MA: MIT Press.

Harris, K. D. 2005. Neural signatures of cell assembly organization. *Nature Reviews Neuroscience*, 6: 399-407.

Hebb, D. O. 1949. *The Organization of Behavior*. New York: Wiley.

Saxton, M. 2010. *Child Language*. London: Sage.

van der Velde, F.; and de Kamps, M. 2006. Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, 29: 37-70.

van der Velde, F.; and de Kamps, M. 2011. Compositional connectionist structures based on in situ grounded representations. *Connection Science*, 23: 97-107.

Autonomy and Intelligence

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

1. Agreements

In general, I agree with the major arguments made in the target article, especially on the following two key points: a truly intelligent system must be able to *learn and adapt*, as well as to manage its resources while working in real time. On this issue, their position is basically the same as mine. In Wang (2006), I define "intelligence" as "the capacity of an information system to adapt to its environment while operating with insufficient knowledge and resources", where

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

"insufficient knowledge and resources" is further specified as having *finite* information-processing capacity, working in *real time*, and *open* to novel tasks. In the book, I argued in detail for why such a working definition is better than the alternatives. These two points still deserve to be stressed, since at the current time most systems get their problem-solving capability *by design*, rather than *from learning*, and the existing learning methods are mostly applicable only to very special situations. Theoretical models of intelligence still routinely ignore the need for the system to work in real time and to manage its own resources, and leave these issues as "implementation details" for the engineers to handle. These attitudes not only may lead the researchers to prefer improper approaches, but also make them to specify the problem incorrectly.

2. Explanations about NARS

One of the AGI systems reviewed in the target article is NARS, my own project. The description of NARS is fairly accurate, though there are some issues to be further clarified. Especially, the review is mostly based on an old version of NARS as described in Wang (2006), while in recent years the new development has addressed most of the issues mentioned in the target article. A new book (in press, Wang 2012), will provide a formal and detailed description of a new version of NARS. In the following I only briefly summarize the relevant parts.

(A) Initial priority of tasks: NARS dynamically distributes its resources (most importantly, processor time) among its reasoning tasks, and the share of time a task gets depends on its priority value. Previously, all input tasks come from a single channel via communication with a human user, who has the option to assign a specific priority value to an input task, otherwise a default value is used. This option is necessary, since it lets the user influence the system's prioritizing of tasks. In the new version, there will be multiple input/output channels between the system and various outside (human or computer) systems and (sensorimotor) devices. Though they will still be given the option to attach a priority value to each task, NARS does not necessarily grant all of them such rights. Instead, the system can decide the initial value for each input task, according to various factors, including the recommended value from the other systems/devices. This change will surely give the system more control of its resources, and at the same time still allow the other systems to have some influence.

(B) Temporal information: Previously in NARS, temporal information was mainly represented by two basic temporal relations (*before-after* and *at-the-same-time*) between events, though absolute time can be represented as special events, and time-related concepts can be explicitly specified. In the new version, this representation is further extended: (1) The system has an internal clock that uses the inference cycle as unit, and (2) the system can use special operations to access outside timing devices. In this way, real-time management of various internal and external processes becomes easier. Even so, it has been argued in Wang (2012) that for a general-purpose AI system, a *relative* representation of time is more fundamental than an *absolute* representation, because it can more flexibly maintain different time scales in different domains, and be more tolerant to the insufficiency of knowledge and resources.

(C) Sensorimotor capacity: Since NARS is usually presented as a reasoning system, it gives many people the impression that it is "symbolic" and "abstract", and cannot easily have sensorimotor capacity. This is not true even for the old version. In Wang (2006) and many other publications, it has been clarified that the "terms" in NARS are not traditional "symbols" that get their meaning by referring to outside objects. Instead, NARS uses an "experience-grounded" semantics, according to which the meaning of a term is determined by its experienced relations with other terms. Perceptual patterns and action schemas can all be represented as terms, and their

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

relations as statements, to be reasoned upon by the inference rules of the system – it is conjectured that the "logic of categorization" is also the "logic of perception" and the "logic of action". Furthermore, in the new version the system has a "sensorimotor interface" that allows a hardware/software device to be connected to NARS and to be used as a "tool" or "organ", which can serve as a sensor, an actuator, or both. In this way, the system can directly get information from its environment and cause outside changes, and all these processes are handled uniformly in the framework of a reasoning system. Here the notion of "reasoning" has been greatly extended to cover cognitive functions including learning, planning, perceiving, and so on.

(D) Meta-learning: As a reasoning system, NARS has a well-defined separation between an object-level (containing acquired knowledge) and a meta-level (containing inference rules and control algorithms). Being *non-axiomatic*, all object-level contents of NARS can be learned and modified. On the contrary, the meta-level content cannot be modified permanently by the system, though in the new version, the system can execute certain "mental operations" to temperately modify its internal status, as well as to learn problem-solving and self-regulating strategies.

3. Disagreements

My major problem with the target article is in its conceptual framework. Some key notions, such as "autonomy", "architecture", and "meta-learning", should be more clearly specified to avoid ambiguity, since they often mean different things in different systems. For example, if an inference rule of system A is emulated by an implication statement in system B, then its modification will be considered as "meta-learning" in A but merely "learning" in B. In this case, to say that A is more autonomous than B sounds unfair. Also, to allow more parts to be modified does not necessarily make a system more intelligent. The rationale and justification of the modifications needs to be evaluated, too. We cannot expect a system to learn everything or to make all design decisions. No matter how autonomous a system is, it needs an invariant core.

References

Wang, P. 2006. *Rigid Flexibility: The Logic of Intelligence*. New York, NY: Springer.

Wang, P. 2012 (in press). *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. Singapore: World Scientific.

Autonomy, Isolation, and Collective Intelligence

Nikolaos Mavridis NIKOLAOS.MAVRIDIS@NYU.EDU
New York University Abu Dhabi
P.O. Box 129188
Abu Dhabi, UAE

Is it total self-sufficiency that we are really after, or harmonious integration into, and facilitation of, ecosystems of intelligent entities, which can participate in wider entities beyond themselves?

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

In the last years, significant progress of numerous partial aspects of cognitive systems has taken place. We now have various theories, as well as laboratory or real-world examples of implemented subsystems, for such partial aspects as: targeting perception, motor control, inference, planning, affect, as well as many flavors of learning. However, a much smaller amount of research has targeted the further integration of these partial aspects. Even less work has been focused on cognitive architectures exhibiting a considerable degree of completeness and generality, which could potentially be applied to a wide variety of application domains, with minimal manual customization or redesign, and which could adapt themselves to changing environments and needs. At the same time, despite the big successes of artificial intelligence and embodied systems in some specific (and usually narrow) domains (Campbell, Hoane and Hsu 2002), the initial big promises as well as the estimated potential of the field have certainly not been fulfilled and reached yet. At the same time many new theoretical as well as practical tools have become available.

Thus, the review paper of Thórisson and Helgasson, appears at a crucial time for the further development of cognitive architectures, and fulfills an important need. Furthermore, it is built around a clear stance, which is applied consistently throughout its exposition. A very short summary of the basic stance of the paper could read like: "The goal of cognitive architectures is to act as a strong basic framework for creating systems with big generality, high autonomy, and strong capabilities for flexible adaptation and deep learning. An ideal example of what we mean by autonomy is a system that can (a) handle a large variety of environments – be it space, desert, ocean floors, and (b) fulfill its goals, which are given only in a high-level description without pre-encoded domain specific knowledge, while (c) operating in isolation. Furthermore, there are four main themes when designing or analyzing cognitive architectures: real-time operation, learning, resource management, and meta-learning. These four dimensions can be used as an analytical framework for describing existing approaches, and also for cross-comparing them, when we quantify their performance across these dimensions".

The authors indeed introduce their stance clearly, as well as follow it consistently throughout the paper; they analyze and cross-compare nine of the most prominent cognitive architectures of today, while providing interesting insights regarding promising avenues for the future. Having had extensive first-hand experience in designing large-scale systems that integrate multiple aspects of cognition, such as the conversational manipulator robot Ripley and its underlying "Grounded Situation Model" architecture (Mavridis and Roy 2006; Mavridis 2007), I am highly sympathetic as well as appreciative of the author's attempt. Furthermore, from the viewpoint of the wider circle of ideas of close proximity to my background, there are a number of interesting observations to be made that could juxtapose to and potentially enrich the review, and provide avenues for future extensions. In more detail:

1) **Autonomy, Isolation, and Collective Intelligence**

In the central stance of the paper which is summarized above, indeed (b) as well as (a) seem to be very good choices for the requirements for ideal autonomous systems. However, requirement (c), namely operation in isolation, needs further examination. Indeed, one can posit this as a requirement of autonomous systems: total self-sufficiency; a machine that, once created, could be left to even operate in a universe where no other intelligent entities exist. The question though follows: is this a good requirement to impose? How productive would that really be, and how relevant to real-world applications, with the exception of harsh environments with no life and big physical difficulties in communications, such as the outer space?

Human intelligence, seen through the lens of effective intelligence, i.e. the capacity for successful action towards self-selected goals, is very much enhanced due to the social nature of

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

humans. And it is not only collaborative teamwork that contributes towards this extension; social learning (Thomaz and Cakmak 2009), learner-directed or observation-based, as well as imitation (Nehaniv and Dautenhahn 2007) and en-culturation, are extremely important for the effective intelligence of humans. Examples of individuals having grown up in isolation provide a strong empirical basis for such an argument (Wikipedia 2012; Davis 1947). Furthermore, one can extend the argument much beyond physically and directly communicating humans. Through oral transmission and writing, the knowledge base of humanity is expanding: and past generations are contributing to our current capacity for effective action. Thus, this is yet another way through which collective intelligence is boosting individual intelligence. Furthermore, we are increasingly entering into frequent interactions with intelligent machines (Mavridis 2011a), which become part of our social networks. But how can this be relevant to cognitive architectures?

2) **Social Competencies and Environments**

To exhibit highly effective intelligence, a system implemented in a cognitive architecture should possess competencies for social interaction, learning and adaptation, which would enable it to utilize not only the physical affordances, but also the social affordances that are available in its environment. Thus, it should be able to participate in human-and-machine social networks (Mavridis 2011b), and position itself in relations with high social capital (Putnam 1993). Ideally, such a system should not only be acting towards its own direct short-term goals, but also contributing towards the increase of the resulting collective effective intelligence of the network that it is participating in. This could be achieved not only through participation in the interactions of the network, but also through facilitating structural rearrangements of the social network around it, for example in order to enrich collective social capital (Social Capital 2012). In that respect, for example part (a) of the basic stance of the authors could be extended to "a wide variety of environments, physical as well as socio-technico-cultural", and certainly (c) "operating in isolation" could be replaced with "operating in sustainable symbiotical interaction with the other intelligent entities that are accessible to it".

3) **Language, Non-verbal Communication, Situation Models, Theory-of-mind**

In order for a system to be able to exhibit the social competencies that were discussed above and participate in human-machine social networks, a basic prerequisite, among others, is to be able to support adequate human-machine as well as machine-machine interaction capabilities. Such interaction capabilities, which need to cover natural language as well as non-verbal communication for the human-machine case, necessitate the existence of real-world solutions to the symbol grounding problem (Harnad 1990), as well as situated and embodied language learning capabilities. Thus, one needs to go much beyond capacities for body and affordance discovery (Saegusa, Metta and Sandini 2010; Stoytchev 2005); and extend to conceptual alignment with other intelligent entities (Goldstone and Rogosky 2002), language learning and social affordance discovery, among others.

But then the question follows: what consequences do such competency requirements have, in terms of cognitive architectures? One possible suggestion here is that, in the same way that the demand for explicit attentional mechanisms arises, explicit situation modeling (Zwaan and Radvansky 1998) representations and standardized processes, such as those implemented in Mavridis and Roy (2006); Mavridis (2007), could be highly useful. Such situation modeling mechanisms can facilitate the bidirectional connection of natural language to the senses; the modeling not only of physical, but also of agentive aspects of the situation. Thus it should support theory-of-mind (Premack and Woodruff 1978) and self, as well as hetero-models; and also

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

directly extend to episodic memories (Mavridis and Petychakis 2010) and predictions. Of course, distinctions between different kinds of memory stores and knowledge bases exist in other cognitive architectures, but rarely there is explicit support for self- and hetero-modeling with theory-of-mind, with the exception of (Friedlander and Franklin 2008). Also, there is very rarely explicit support for the transition between natural language and symbolic representations, and for on-the-fly conceptual and situation-model alignment across agents, which would be highly valuable.

4) **Embodiment, Collective Intelligence and Offloading to Distributed Services**

Furthermore, as thousands of services are becoming available through the internet, in order for a system to take advantage of the full capabilities of the human-machine social network and extend beyond the limitations of its own physical embodiment and processing faculties, it should be able to interact with and utilize remote sensing, actuation, processing, and storage resources. Such services could either be provided by machines (such as those available on a computation or sensing cloud (Amherst, Fox, et al. 2010) or even by humans (such as the real time visual sensing and recognition services provided by humans in the DARPA Network Challenge (10 Red Balloons). Thus, for example, much before the limitations of computational power available by an onboard CPU are surpassed, such a system can harvest the much greater networked processing power of the cloud. Before tasks which are hard for AI but easy for humans become within reach of AI, such system can offload these tasks to networked humans which are part-time crowd-servicing (David 2011). In all those cases where a particular machine sensor or actuator – for example a camera – is not available but a human is, it can utilize the human services to achieve its goal, as in the 10 Red Balloons challenge, and effectively act as if the human sensing subsystems were temporarily part of its own embodiment. This is the idea advocated in the Human-Robot Cloud (Mavridis 2012), which enables the on-the-fly construction and reconstruction of distributed human-machine cognitive systems.

Taking the above four points into account, one resulting extension of the basic stance could be summarized as: "Real-time operation, Resource Management, Learning, and Meta-Learning – but beyond the limitations of the individual system: extending towards the total resources and total embodiment of the human-machine social network of which it is a part. Thus, the system is actually viewing the network as if the sensing, actuation, and processing services offered through it are part of its own body and resources. Starting from within the viewpoint of the individual entity, it needs extending to a more holistic consideration of the resources of the network of which it is part of, manages these resources, and does not only participate with real time considerations, but can also actively maintain and shape them. Also structurally – effectively performing network-wide learning – it can participate in a network-wide self-reflection and meta-learning mechanisms. In this way we extend the four main themes of the review paper from the individual intelligent entity (its own physical and informational resources) to the whole network, but always from within the viewing capabilities of the individual entity, as it pertains to its view of the human-machine network of which it is part of.

In summary: instead of connecting "autonomy" with a requirement of total self-sufficiency and capability of operation in isolation, it is much more reasonable to center our efforts towards positioning the intelligent entities created through cognitive architectures appropriately within human-machine social networks, externally offloading their physical and informational functions when needed, and harmoniously and empathetically integrating within ecosystems of intelligent entities, so that they can participate in much wider entities beyond themselves.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

References

Armbrust, M.; Fox, A.; Griffith R; et al. 2010. A View of Cloud Computing. *Communications of ACM*. 53: 50-58.

Campbell, M.,; Hoane, A. J.; and Hsu, F. 2002. Deep Blue. *Artificial Intelligence*. 134: 57-83.

Davis, J. G. 2011. From Crowdsourcing to Crowdservicing. *IEEE Internet Computing*. 15: 92-94.

Davis, K. 1947. Final Note on A Case of Extreme Isolation. *American Journal of Sociology*. 52: 432-437.

Goldstone, R. L.; and Rogosky, B. J. 2002. Using Relations within Conceptual Systems to Translate across Conceptual Systems. *Cognition*. 84: 295-320.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D*. 42: 335-346.

Mavridis, N. 2007. *Grounded Situation Models for Situated Conversational Assistants*. Ph.D. diss., MIT.

Mavridis, N. 2011a Growing Robots in the Desert", TEDx Al Ain video available electronically at http://www.youtube.com/watch?v=HNSqQKZiofA

Mavridis, N. 2011b. Artificial Agents Entering Social Networks. In *A Networked Self*, Routledge *291-303*.

Mavridis, N.; Bourlai, T.; and Ognibene, D. 2012. The Human-Robot Cloud: Situated Collective Intelligence on Demand. In *Proceedings of IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems*, 360-365.

Mavridis, N.; and Petychakis, M. 2010. Human-like Memory Systems for Interactive Robots: Desiderata and Two Case Studies Utilizing Grounded Situation Models and Online Social Networking. In *Proceedings of 36th Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*.

Mavridis, N.; and Roy, D. 2006. Grounded Situation Models for Robots: Where Words and Percepts Meet. In *Proceedings of IEEE IROS*, 4694-4697.

Nehaniv, C. L.; and Dautenhahn, K. 2007. *Imitation and Social Learning in Robots, Humans and Animals*. Cambridge University Press.

Premack, D.; and Woodruff, G. 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences* 1: 515-526.

Putnam, R. D. 1993. The Prosperous Community: Social Capital and Public Life. *The American Prospect* 13: 35-42.

Saegusa, R.; Metta, G.; and Sandini, G. 2010. Self-Body Discovery Based on Visuomotor Coherence. In *Proceedings of 3rd Conference on Human System Interactions*, 356-362.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

Stoytchev, A. 2005. Toward Learning the Binding Affordances of Objects: A Behavior-Grounded Approach. In *Proceedings of AAAI Symposium on Developmental Robotics*, 17-22.

Social Capital, 2012. Available electronically at: http://www.socialcapitalresearch.com/definition.html

Thomaz, A. L.; and Cakmak, M. 2009. Social Learning Mechanisms for Robots. In *Proceedings of International Symposium on Robotics Research* 1-14.

Wikipedia, 2012. Humans Growing in Isolation. Available electronically at: http://en.wikipedia.org/wiki/Feral_child

Zwaan, R. A.; and Radvansky, G. A. 1998. Situation Models in Language Comprehension and Memory. *Psychological Bulletin*. 123: 162-185.

10 Red Balloons, DARPA Network Challenge 2009. Available electronically at: http://en.wikipedia.org/wiki/DARPA_Network _Challenge

Response to Comments

Kristinn R. Thórisson[1,2] THORISSON@RU.IS

Helgi Páll Helgasson[1] HELGIH09@RU.IS

[1]*Center for Analysis & Design of Intelligent Agents / School of Computer Science*
Venus, 2nd fl.
Reykjavik University
Menntavegur 1, 101 Reykjavik, Iceland

[2]*Icelandic Institute for Intelligent Machines*
2. h. Uranus
Menntavegur 1, 101 Reykjavik, Iceland

We thank the reviewers for the commentary on our paper, and their numerous suggestions for expansion; we appreciate the opportunity to address some of our main points from other angles. As there are few overlaps between the reviews we give here a separate reply to each. Our motivation to write the paper stems from important but much-ignored issues in architecture design and methodology; rather than being lead in the various directions that the reviews invite us to do, our approach has been to highlight some of the key points we make in the paper, as relevant in each case, because we really think these must be addressed by the AGI community at large. None of the reviewers challenge the core points or main conclusions in our paper, but offer suggestions for improvements and

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

extensions to the way it is presently written. To us this implies the need for one or more follow-up papers, but these may not necessary be written by us. It is also clear, after reading the reviews carefully, that not all of the reviewers are convinced of the importance of the topic of autonomy and how deeply the implications run for AGI architectures. We may find that those who remain unconvinced are researchers whose primary research goals lie outside of AGI; we certainly hope so, because we really want all AGI researchers to take a serious look at issues of autonomy.

1. Cristiano Castelfranchi, ISTC-CNR

We appreciate that others have devoted quite some time to thinking about and debating what autonomy is. In our paper we are focusing on a particular interpretation of autonomy that is intended to be in line with a generally-accepted definition that one might expect a large subset of society to agree with; we are not interested in defining autonomy per se, or to get into the deeper arguments for what does or does not constitute autonomy. This is why we provide an illustrative example of what kind of system we envision when we use the term autonomy. We would be surprised to see disagreement amongst the general population about whether our autonomous robot example is on the border of what one would define as "autonomous" – we think it is very clear indeed that it would be. Thus, while we do not provide a formal theory of autonomy for AGI systems, the meaning of autonomy in the context of the paper is articulated beyond vagueness.

Figure 1 presents an autonomy comparison framework, explicitly showing several components and dimensions of autonomy. Given Castelfranchi's classification of autonomy, one kind of autonomy we consider in our paper might be mapped to his "goal autonomy" – systems that are autonomous from the designer/operator point of view. In a next iteration of our discussion we might well delve into distinctions between "physical environment" and "social environment", but at the present (relatively primitive) state of cognitive architectures the dimensions that we use are quite sufficient, and finer distinctions such as those Castelfranchi suggests, seems to us a bit over the top and possibly distracting away from the main point of the paper, namely to do a first approximation of the ability of current architectures to address autonomy, in the reasonably intuitive sense of that term.

2. Antonio Chella, *DICGIM – University of Palermo*

Our autonomy comparison framework presented in Figure 1 touches on embodiment, but this is the least of our interests. While the level of autonomy achieved in a (e.g. robotic) agent depends of course on its embodiment, a key ignored aspect of present architectures is their potential to achieve the kinds of autonomy one would anticipate an artificially generally intelligent agent to need. This is the core issue addressed in the paper. On a side note, we see no reason to believe that a particular kind of body or environment is needed to achieve intelligence, granted that more complex environments will require higher

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

levels of intelligence, and so far no virtual environment exists that displays comparable levels of complexity as the real world. Developmental and causal processes between the agent and its environment are not purely embodiment issues, these rely both on embodiment and other cognitive processes (e.g. learning) stated as part of the autonomy comparison framework.

3. Vassilis Cutsuridis, *Kings College London*

In viewing processes such as resource management, real-time processing, learning, and meta-learning as "embodiment" functions, as done by Cutsuridis, fundamental cognitive processes are being misrepresented as seemingly optional to artificial general intelligence (AGI) architectures. While we have great difficulties envisioning an intelligent system without some form of embodiment, it should be clear to anyone reading our paper that we fully appreciate the relevance of other cognitive functions (e.g. memory, perception, reasoning and planning); nowhere do we suggest they are less important to AI systems than the ones we focus on in this paper. Neither we, nor anyone else for that matter, knows the complete list of necessary and sufficient functions for achieving artificial general intelligence. In this context we must keep in mind that when attempting to build an artificial general intelligence a large number of cognitive functions are necessary – but not sufficient. Autonomy, broken down into the necessary (but not sufficient) functions of real-time processing, resource management, learning, and meta-learning, is in our opinion necessary. Nothing in Cutsuridis's comments seems to contradict that.

We generally agree with Pei Wang's definition of intelligence (Wang 2006), which inevitably leads us to focus on real-time processing and resource management. Central to intelligence, then, is the interaction of the system with the environment. While Cutsuridis believes that "starting with embodiment is like building a house from the roof down", we disagree and find the opposite to be true. Designing an AGI system possessing a range of different cognitive functions without any thought of practical operation (which is perhaps more in line with what Cutsuridis refers to as embodiment than our understanding of that term) until the final stages can have potentially catastrophic results. Due to the tight coupling of cognitive functions, it may be extremely difficult to retrofit such a system with resource management and real-time processing capabilities. In our view, time and resource constraints are an integral part of intelligence and must be considered at all levels.

It behooves us to correct a couple of other misunderstandings as well. Firstly, we do not, as the authors imply, consider perception-action capabilities somehow "less important" than autonomy. We believe that to create an architecture and system capable of more than narrow artificial intelligence – that is, an artificial general intelligence (AGI) – a number of issues must be addressed. We believe autonomy has fallen by the wayside in much of the past and present research on AGI, and our paper was written to highlight the important aspects of cognition that surface when we look more closely at this issue. Secondly, Cutsuridis discussion turns to the issue of reasoning. While we do not fully share his view that reasoning is the "highest faculty" of the human brain, we agree that it is an important – and quite possibly a critical – aspect of cognition. However,

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

with regards to autonomy, "reasoning" is too coarse-grain a concept to help us understand what present architectures might be missing to achieve it. Thirdly, the authors argue that none of the architectures reviewed in our paper are "brain-inspired". In our view all cognitive architectures are "brain-inspired" at some level of detail. The authors' comments seem to imply that "brain-inspiration" is a special target of ours, which is incorrect, although we do believe inspiration from biology and cognitive science is a good thing for AI. However, that is not the subject of our paper. As clearly stated throughout the paper, we are interested in assessing the ability of current architectures – those with serious aspirations towards AGI – to address the various aspects of autonomy that we find to be necessary (but not sufficient) for achieving that goal.

4. Włodzisław Duch, *NCU Toruń* & *NTU Singapore*

We share Duch's view that creativity is an important aspect of AGI systems. Learning implies that a system will perform new behaviors, as it increases its own knowledge, to solve new problems or solve old ones more efficiently. As the power of a system's learning mechanisms increases it should be capable of generating more sophisticated new behaviors. This may be viewed as a way of achieving creativity, which arises as an emergent surface phenomenon as a result of the cognitive processes of the underlying architecture. Striving directly for creativity for its own sake does not seem to be a fruitful approach to us. To us, creativity is more of a measure of the learning power of the system rather than an explicit cognitive function. In its most watered-down meaning, all animals that can generate new behavior (given some measure of minimum difference to justify calling it "new") can be said to be creative. A cognitive agent that is capable of generating truly novel, non-obvious solutions to problems is more creative than one that solves problems with run-of-the-mill solutions. However, only at the low end of this spectrum, i.e. given the watered down definition, does creativity seem necessary for achieving autonomy.

5. Ryan McCall & Stan Franklin, *Fedex Institute of Technology*

McCall & Franklin point out that no formal definition of meta-learning is provided in our paper. This is true, but that would be a major undertaking and could easily take up the space of several papers, taking the reader down a different path. The utility of a formal definition of meta-learning is in any case not obvious, and our focus is somewhat at higher-level. We clearly articulate meta-learning as the general process of improving one's own learning performance and capability mechanisms, which is a sufficiently specific description for the purposes of our comparison and evaluation. Biological agents (e.g. humans) are limited in the dimension of meta-learning by fixed operating principles and structures of the brain. This precludes the possibility of these agents performing any kind of meta-learning requiring functional change through deliberate structural modification at the architecture level. However, the level of structural change possible in software systems is limited primarily by their design. A properly designed software

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

system can exhibit a high degree of plasticity and flexibility, even at the structural level.

Upcoming results from the AERA architecture (which were not ready to be presented in this paper) show a learning agent that changes its own operation and structure at run-time. Of course random changes to the inner workings of an agent are likely to cause catastrophic failures, however the same does not apply to directed, pre-contemplated changes, the effects of which may already have been simulated, predicated and evaluated beforehand, as demonstrated in AERA.

McCall & Franklin ask why meta-learning is a research priority when a working human level AGI system has not yet been produced. Our answer is along the lines of what we argue in the paper, namely that we are not optimistic that human-level AGI can be achieved (any time soon) without meta-learning. Due to the complexity inherent in such a system, the cognitive limits of human designers/programmers, as well as the lack of suitable software development methodologies for such large and complex systems, we see no other way than to delegate responsibility for growth and development of such a system onto the system itself – manual implementation does not appear to be an option. This becomes particularly clear when considering the largest, most complex existing software systems today and compare them to the types of systems likely to be needed for human level AGI. This is why we make self-reconfiguring systems a research priority. Self-directed growth should be in the direction of greater performance and capability, closely related to meta-learning.

Finally, we appreciate the clarifications of the LIDA architecture.

6. Brandon Rohrer, *Sandia National Laboratories*

Rohrer points out the very important issue of having clear design and research goals, when building AGI architectures, a point we absolutely agree with. He further emphasizes that the explicit goal of a number of cognitive architectures is to model the operation of the human brain, rather than being driven by ambitions of practical AGI systems performing useful tasks in the world. Our paper sets a very practical frame to the discussion, with an exploration robot scenario at the center – this illustrative example, and the general discussion throughout the paper, makes it very clear that our target is of a practical nature. So our paper targets those interested in creating useful AGI systems – systems with practical application. While we do not believe that such a system, should one be created in the near future, would have zero explanatory power in regard to the operation of the brain, we understand that pursuing the goal of explaining the operation of the brain will often lead down very different paths. But our experience tells us – so far – that if AGI requires deep understanding of architecture, looking at the brain is not the best place to go: a lack of the big picture is exactly what seems missing from all brain research to date. The pursuit of architectural principles for AGI along alternative avenues seems to us more likely to bear fruit.

One of the goals of this article is highlighting topics which, while having been largely ignored, are nevertheless keys to achieving the “G” in “AGI”. As we argue that architecture cannot be ignored, we thus fully agree with Rohrer that clear and explicit research goals and evaluation methods are important to the field, reducing confusion and

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

focusing scientific discussion more efficiently.

7. Frank van der Velde, *University of Twente*

Our vision coincides with the key message of Van der Velde's review, that cognitive growth is key to general intelligence, and that for such growth to succeed a cognitive architecture must be sufficiently exposed to environments conductive to such growth. He criticizes our analysis of autonomy for being "cognitively naïve", as our exploration robot example describes operational capacity not necessarily exhibited by human beings. Note that our paper is not titled "How to design AGI architectures: The complete guide". This purported cognitive naïvité is thus by design. We are highlighting one aspect of AGI, autonomy, that has fallen too much by the wayside, in our opinion, yet has vast implications for how they are designed, as we subsequently proceed to demonstrate. Make no mistake, autonomy is only one aspect of many that are necessary for achieving AGIs. Another one is precisely what der Velde's review mostly revolves around, cognitive growth. It should be noted that we use the term "learning" more inclusively than he does, and we see cognitive growth as a necessary aspect of being able to learn a vast array of tasks and adapt to and operate in many contexts and domains. Learning may occur at different levels and we would probably consider what Van der Velde refers to as "development" to be structural learning, i.e. changes to the structure of the cognitive system, based on experience. To clarify a possible misunderstanding, we do not view learning as a process of learning method selection.

Some of Van der Velde's comments concern preparation for various environments. It is true that human explorers are unlikely to venture into very exotic environments without special preparation; otherwise they would not survive, at least not in the outer space. Our example of the exploration robot is of course a hypothetical one, constructed to highlight aspects of cognitive capabilities. In the paper we explicitly say that "the robot … is designed to physically withstand the ambient environmental conditions of these environments". This gives the robot greater opportunity for safely learning "on the job" than humans would have. We would also argue that high plasticity is not only beneficial at the early stages of operation for a cognitive architecture. While the brain can be said to lose its plasticity in adulthood, this is primarily a biological limitation we do not see as necessary or always useful for AI systems. AI systems could also usefully be designed to exhibit a greater degree of plasticity during earlier stages of operation than in later stages. We see no reason to limit plasticity at later stages; for some uses of AGI systems the potential for plasticity could remain useful. Should the system for example face changes in the environment or tasks at later stages, plasticity will become more useful (and perhaps necessary) again.

8. Pei Wang, *Temple University*

We agree with Pei Wang that the degree of self-modification possessed by a system alone does not directly relate to its degree of intelligence, but then again here we are not

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

discussing self-modification in light of higher or lower intelligence, but rather higher or lower levels of autonomy. The list of cognitive functions we have deemed necessarily implied from the requirement of autonomy establishes a minimum list of necessary (but not sufficient) processes for achieving AGI. The power of meta-cognitive processes in a system influence how much the intelligence of the system can be increased by using self-modification. So a system capable of self-modification has more intelligence potential in a sense, compared to a static system incapable of self–modification. An "invariant core", at least in the form of some low-level operating principles, is needed for all AI systems, regardless of their levels of autonomy, if we want the system to have some level of predictability.

Further clarification of the concepts of autonomy, architecture and meta-learning will be beneficial – but might be worthy of dedicated papers rather than having been included in ours. We appreciate the explanations with regards to NARS and its intended future directions, and look forward to demonstrations of the increased capabilities resulting from the improvements.

9. Nikolaos Mavridis, *New York University, Abu Dhabi*

Mavridis argues that making isolated operation a requirement for autonomous AI systems may not be appropriate as there may be substantial benefit in focusing on social aspects (system-to-system or system-to-human interaction), collective intelligence (systems sharing their knowledge) and cloud computing (distributed resources). We agree that these aspects are highly interesting and should absolutely be pursued in the future. However, this view does not contradict but rather extends the points we make in the paper, that autonomy needs to be addressed in architectures, and that present ones are not in a good position to support it. Since it is not clear, at this point, how requirements for autonomy may be relaxed when taking Mavridis' stance, the fact remains that if any autonomy is to be attained at all, current efforts fall short.

To elaborate further, self-sufficiency is a requirement for full autonomy in the sense that a system should make its own goals. Accepting goals from other social entities is not precluded by this, but the system should make its own decision whether to accept such a goal in each case. When we take a general approach to intelligence, as done in the field of AGI, social interaction becomes a domain for systems that should be, by definition, as domain-independent as possible; any specific domain requires domain-specific skills. Social interaction may be one of several domains a single system must simultaneously operate in; a household robot responsible for performing chores and being capable of natural language communication with people would be an example of such a system. However, we do not believe that social interaction is fundamental for intelligence in the most general sense. While we list learning-by-observation as an example of learning style in Figure 1 of our paper, this is meant in a general sense and does not necessitate the presence of another intelligent entity, while we concede that there is more to be learned from an environment containing intelligent entities than an environment devoid of such. The use of collective intelligence as an environment for cognitive growth and learning does not contradict what we discuss in our paper, although our example of the exploration

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM

robot does not mention it.

Conceptually, it makes little difference whether the resources of the system are physically confined or grouped together. For example, the capability to seek knowledge from the Internet is a skill that is not fundamentally different from having a robot turn its head to see what lies in a new direction. In both cases, the system has a goal of gathering information and is able to satisfy that goal with a particular skill, be it motor-control or Internet search. While using distributed resources may give the smallest robot access to vastly more powerful computational resources (processing and memory) than would otherwise be possible, this does not make any dimension in our autonomy comparison framework redundant. Unless all the proposed autonomy dimensions we have suggested are addressed, we have great difficulty envisioning how an embodied AI system possessing general learning capabilities and operating in real-world environments could be constructed, regardless of target domains. However, once such systems can be constructed, the directions proposed by Mavridis are highly interesting, practical and natural ones to follow.

References

Wang, P. 2006. *Rigid Flexibility: The Logic of Intelligence*. New York, NY: Springer.

Brought to you by | Sandia National Laboratory
Authenticated | 198.102.153.2
Download Date | 4/10/13 8:45 PM