

# The architecture challenge: future artificial intelligent systems will require sophisticated architectures, and knowledge on brain might guide their construction

Gianluca Baldassarre, Vieri Giuliano Santucci, Emilio Cartoni, and Daniele Caligiore

*Laboratory of Computational Embodied Neuroscience, Institute of Cognitive Sciences and Technologies, National Research Council of Italy  
Via San Martino della Battaglia 44, 00185 Roma, Italy  
{gianluca.baldassarre,vieri.santucci,emilio.cartonidaniele.caligiore}@istc.cnr.it*

## Abstract

In this commentary, we highlight a crucial challenge posed by the proposal of Lake et al. to introduce key elements of human cognition into deep neural networks and future artificial intelligence systems: the need to design effective sophisticated architectures. We propose that looking at the brain is an important means to face this great challenge.

## Commentary

We agree with the claim of Lake et al. that to obtain human-level learning speed and cognitive flexibility, future artificial intelligence (AI) systems will have to incorporate key elements of human cognition: from causal models of the world, to intuitive psychological theories, compositionality, and knowledge transfer. However, the authors largely overlook the importance of a major challenge to implement the functions they advocate: the need to develop sophisticated *architectures* to learn, represent, and process the knowledge related to those functions. Here we call this *the architecture challenge*. In this commentary, we make two claims: (1) tackling the *architecture challenge* is fundamental to succeeding in developing human-level AI systems; (2) looking at the brain can furnish important insights on how to face the *architecture challenge*.

The difficulty of the *architecture challenge* stems from the fact that the space of the architectures needed to implement the several functions advocated by Lake et al. *is huge*. The authors get close to this problem when they recognize that one main thing that the enormous genetic algorithm of evolution has done in millions of years of stochastic hill-climbing search, is to develop suitable brain architectures. One possible way to attack the *architecture challenge*, also mentioned by Lake et al., would be to use evolutionary techniques mimicking evolution. We think that at the moment this strategy is out of reach, given the “ocean-like” size of the search space. At most, we can use such techniques to explore small, interesting “islands lost within the ocean”. But how to find those islands in the first place? We propose to look at the architecture of real brains, the product of the evolution genetic algorithm, and try to “steal insights” from nature. Indeed, we think that *much of the intelligence of the brain resides in its architecture*. Obviously, identifying the proper insights is not easy to do, as the brain is very difficult to understand. However, it might be useful to try, as the effort might give us at least some general indications, a compass, to find the islands in the ocean. Here we present some examples to support our intuition.

When building architectures of AI systems, even when following cognitive science indications (e.g., Franklin, 2007), the tendency is to “divide and conquer”; that is, to list the needed high-level functions, implement a module for each of them, and suitably interface the modules. However, the organization of the brain can be understood not only on the basis of high-level functions (see below), but also as based on “low level” functions (usually called “mechanisms”). An example of mechanisms is the brain organisation based on macro structures each having fine repeated micro-architectures implementing specific computations and learning processes (Doya, 1999; Caligiore et al., 2016): The cortex to statically and dynamically store knowledge acquired by associative learning processes (Shadmehr & Krakauer, 2008; Penhune & Steele, 2012), the basal-ganglia to learn to select information by reinforcement learning (Houk et al., 1995; Graybiel, 2005), the cerebellum to implement fast time-scale computations possibly acquired with supervised learning (Wolpert et al., 1998; Kawato et al., 2011), and the limbic brain structures interfacing the brain to the body and generating motivations, emotions, and the value of things (Mogenson et al., 1980; Mirolli et al., 2010). Each of these mechanisms supports multiple, high-level functions (see below).

Brain architecture is also forged by the fact that natural intelligence is strongly *embodied* and *situated* (an aspect not much stressed by Lake et al.), i.e. shaped to adaptively interact with the physical world (Anderson, 2003; Pfeifer & Gómez, 2009) to satisfy the organism's needs and goals (Mannella et al., 2013). Thus, the cortex is organised along multiple cortical pathways running from sensors to actuators (Baldassarre et al., 2013) and “intercepted” by the basal ganglia selective processes in their last part closer to action (Mannella & Baldassarre, 2015). These pathways are organised in a hierarchical fashion, with the higher ones processing needs and motivational information, controlling the lower ones closer to sensation/action. The lowest pathways dynamically connect musculoskeletal body proprioception with primary motor areas (Churchland et al., 2012). Higher-level “dorsal” pathways control the lowest pathways by processing visual/auditory information used to interact with the environment (Scott, 2004). Even higher-level “ventral” pathways inform the brain on the identity and nature of resources in the environment to support decisions on what to do (Milner & Goodale, 2006; Caligiore et al., 2010). At the hierarchy apex, the limbic brain supports goal selection based on visceral, social, and other types of needs/goals. Embedded within the higher pathways, an important structure involving basal ganglia-cortical loops learns and implements stimulus-response habitual behaviours (used to act in familiar situations) and goal-directed behaviours (important for problem solving and planning when new challenges are encountered; Mannella et al., 2013; Baldassarre et al., 2013a). These brain structures form a sophisticated network, whose knowledge might help designing the architectures of human-like embodied AI systems able to act in the real world.

A last example of the need for sophisticated architectures starts from the recognition by Lake et al. that we need to endow AI systems with a “developmental start-up software”. In this respect, together with other authors (e.g., Weng et al., 2001; see Baldassarre et al., 2013a, and Baldassarre et al., 2014, for collections of works) we believe that human-level intelligence can be achieved only through *open-ended learning*, i.e. the cumulative learning of progressively more complex skills and knowledge, driven by *intrinsic motivations*, which are motivations related to the acquisition of knowledge and skills rather than material resources (Baldassarre, 2011). The brain (e.g., Lisman & Grace, 2005; Redgrave & Gurney, 2006) and computational theories and models (e.g., Baldassarre & Mirolli, 2013a; Baldassarre et al., 2014; Santucci et al., 2016) indicates how the implementation of these processes indeed requires very sophisticated architectures able to store multiple skills, to

transfer knowledge while avoiding catastrophic interference, to explore the environment based on the acquired skills, to self-generate goals/tasks, and to focus on goals that ensure a maximum knowledge gain.

## References

- Anderson, M. L. (2003). Embodied Cognition: A field guide. *Artificial Intelligence* 149(1):91-130.
- Baldassarre, G. (2011). What are intrinsic motivations? A biological perspective. In: Cangelosi, A., Triesch, J., Fasel, I., Rohlfing, K., Nori, F., Oudeyer, P.-Y., Schlesinger, M. & Nagai, Y. (ed.), *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*. E1-8. New York, NY: IEEE.
- Baldassarre, G. & Mirolli, M. (eds.) (2013a), *Intrinsically motivated learning in natural and artificial systems*. Springer, Berlin.
- Baldassarre, G., Caligiore, D. & Mannella, F. (2013). The hierarchical organisation of cortical and basal-ganglia systems: a computationally-informed review and integrated hypothesis. In: Baldassarre, G. & Mirolli, M. (eds.) (2013), *Computational and Robotic Models of the Hierarchical Organisation of Behaviour*. 237-70. Springer-Verlag.
- Baldassarre, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K. & Mirolli, M. (2013a). Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model. *Neural Networks* 41: 168-87.
- Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M. & Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Frontiers in psychology* 5.
- Caligiore, D., Pezzulo, G., Baldassarre, G., Bostan, A. C., Strick, P. L., Doya, K., Helmich, R. C., Dirks, M., Houk, J., Jörntell, H., Lago-Rodriguez, A., Galea, J. M., Miall, R. C., Popa, T., Kishore, A., Verschure, P. F. M. J., Zucca, R. & Herreros, I. (2016). Consensus paper: Towards a systems-level view of cerebellar function: The interplay between cerebellum, basal ganglia, and cortex. *The Cerebellum* 1-27. (doi: 10.1007/s12311-016-0763-3).
- Caligiore, D., Borghi, A., Parisi, D. & Baldassarre, G. (2010). TRoPICALS: A computational embodied neuroscience model of compatibility effects. *Psychological Review* 117 (4):1188-1228.
- Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I. & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature* 487:51-56.
- Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks* 12 (7-8):961-74.
- Franklin, S. (2007). A foundational architecture for artificial general intelligence. In: Want, P. & Goertzel, B. (ed.), *Advances in artificial general intelligence: Concepts, architectures and algorithms*. 6:36. Google Books.
- Graybiel, A. M. (2005). The basal ganglia: learning new tricks and loving it. *Current Opinion in Neurobiology* 15(6):638-44.
- Houk, J. C., Adams, J. L. & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Houk, J. C.; Davids, J. L. & Beiser, D. G. (ed.), *Models of Information Processing in the Basal Ganglia*. 249-70. MIT Press.

- Kawato, M., Kuroda, S. & Schweighofer, N. (2011). Cerebellar supervised learning revisited: biophysical modeling and degrees-of-freedom control. *Current Opinion in Neurobiology* 21(5):791-800.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. (2017). Building machines that learn and think like people. *Brain and Behavioural Sciences*.
- Lisman, J. E. & Grace, A. A. (2005). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron* 46:703-13.
- Mannella, F. & Baldassarre, G. (2015). Selection of cortical dynamics for motor behaviour by the basal ganglia. *Biological Cybernetics* 109:575-95.
- Mannella, F., Gurney, K. & Baldassarre, G. (2013). The nucleus accumbens as a nexus between values and goals in goal-directed behavior: A review and a new hypothesis. *Frontiers in Behavioral Neuroscience* 7(135):e1-29.
- Milner, D. & Goodale, M. (2006). *The visual brain in action*. Oxford University Press.
- Mirolli, M.; Mannella, F. & Baldassarre, G. (2010). The roles of the amygdala in the affective regulation of body, brain and behaviour. *Connection Science* 22(3):215- 45.
- Mogenson, G. J., Jones, D. L. & Yim, C. Y. (1980). From motivation to action: Functional interface between the limbic system and the motor system. *Progress in Neurobiology* 14 (2-3):69-97.
- Penhune, V. B. & Steele, C. J. (2012). Parallel contributions of cerebellar, striatal and M1 mechanisms to motor sequence learning. *Behavioural Brain Research* 226(2):579-91.
- Pfeifer, R. & Gómez, G. (2009). Morphological computation—connecting brain, body, and environment. In: *Creating brain-like intelligence* 66-83. Springer Berlin Heidelberg.
- Redgrave, P. & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience* 7:967-75.
- Santucci, V. G., Baldassarre, G. & Mirolli, M. (2016), GRAIL: A goal-discovering robotic architecture for intrinsically-motivated learning, *IEEE Transactions on Cognitive and Developmental Systems* 8(3):214-31.
- Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience* 5(7):532-46.
- Shadmehr, R. & Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Experimental brain research*. 185(3), 359-381.
- Wolpert, D. M., Miall, R. C. & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Science* 2(9):338-47.
- Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M. & Thelen, E. (2001). Autonomous mental development by robots and animals. *Science* 599-600.