Biologically Inspired Cognitive Architectures (2015) **14**, 97–110

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bica

INVITED ARTICLE

Towards integrated neural–symbolic systems for human-level AI: Two research programs helping to bridge the gaps

Tarek R. Besold *, **Kai-Uwe Kühnberger**

Institute of Cognitive Science, University of Osnabrück, D-49069 Osnabrück, Germany

Received 9 September 2015; received in revised form 13 September 2015; accepted 13 September 2015

KEYWORDS
Research program;
Neural–symbolic integration;
Complexity theory;
Cognitive architectures;
Agent architectures

Abstract After a human-level AI-oriented overview of the status quo in neural–symbolic integration, two research programs aiming at overcoming long-standing challenges in the field are suggested to the community: The first program targets a better understanding of foundational differences and relationships on the level of computational complexity between symbolic and subsymbolic computation and representation, potentially providing explanations for the empirical differences between the paradigms in application scenarios and a foothold for subsequent attempts at overcoming these. The second program suggests a new approach and computational architecture for the cognitively-inspired anchoring of an agent's learning, knowledge formation, and higher reasoning abilities in real-world interactions through a closed neural–symbolic acting/sensing–processing–reasoning cycle, potentially providing new foundations for future agent architectures, multi-agent systems, robotics, and cognitive systems and facilitating a deeper understanding of the development and interaction in human-technological settings.

A tale of symbols and signals: the quest for neural–symbolic integration

> "*I repeat my belief that learning has to be at the center of the artificial intelligence enterprise. While I do not regard intelligence as a unitary phenomenon, I do believe that the problem of reasoning from learned data is a central aspect of it.*" Leslie Valiant, Valiant (2013), p. 163

A seamless coupling between learning and reasoning is commonly taken as basis for intelligence in humans and, in close analogy, also for the biologically-inspired (re-)creation of human-level intelligence with computational means. Still, one of the unsolved methodological core issues in human-level AI, cognitive systems modelling, and

---

* Corresponding author.
E-mail address: tarek.besold@uni-osnabrueck.de (T.R. Besold).

http://dx.doi.org/10.1016/j.bica.2015.09.003
2212-683X/

cognitive and computational neuroscience–and as such one of the major obstacles towards solving the Biologically Inspired Cognitive Architectures (BICA) challenge (Samsonovich, 2012)–is the question for the integration between connectionist subsymbolic (i.e., ''neural-level'') and logic-based symbolic (i.e., ''cognitive-level'') approaches to representation, computation (mostly subsymbolic) learning, and (mostly symbolic) higher-level reasoning.

AI researchers working on the modelling or (re-)creation of human cognition and intelligence, and cognitive neuroscientists trying to understand the neural basis for human cognition, have for years been interested in the nature of brain-computation in general (see, e.g., Adolphs, 2015) and the relation between subsymbolic/neural and symbolic/cognitive modes of representation and computation in particular (see, e.g., Dinsmore, 1992). The brain has a neural structure which operates on the basis of low-level processing of perceptual signals, but cognition also exhibits the capability to efficiently perform abstract reasoning and symbol processing; in fact, processes of the latter type seem to form the conceptual cornerstones for thinking, decision-making, and other (also directly behavior-relevant) mental activities (see, e.g., Fodor & Pylyshyn, 1988).

Building on these observations–and taking into account that hybrid systems loosely combining symbolic and subsymbolic modules into one architecture turned out to be insufficient for the purpose–agreement on the need for fully integrated neural—cognitive processing has emerged (see, e.g., Bader & Hitzler, 2005; d'Avila Garcez et al., 2015). This has several reasons also beyond the analogy to the described functioning principles of the brain:

- In general, network-based approaches possess a higher degree of biological motivation than symbol-based approaches, also outmatching the latter in terms of learning capacities, robust fault-tolerant processing, and generalization to similar input. Also, in AI applications they often enable flexible tools (e.g., for discovering and processing the internal structure of possibly large data sets) and efficient signal-processing models (which are biologically plausible and optimally suited for a wide range of applications).
- Symbolic representations are generally superior in terms of their interpretability, the possibilities of direct control and coding, and the extraction of knowledge when compared to their (in many ways still black box-like) connectionist counterparts.[1]
- From a cognitive modelling point of view, subsymbolic representations for tasks requiring symbolic high-level reasoning might help solving, among many others, the problem with ''too large'' logical (epistemic) models (see, e.g., Gierasimczuk & Szymanik, 2011) which seem to lead to implausible computations from the reasoning agent's perspective (Degremont, Kurzen, & Szymanik, 2014). On the other hand, being able to lift subsymbolic brain-inspired models and corresponding simulations to a symbolic level of description and analysis promises to close the interpretative and explanatory gap between actual biologically-motivated model dynamics and observed behavior also for tasks involving complex or abstract reasoning.

In summary, cognitive-level interpretations of artificial neural network (ANN) architectures and accurate and feasible neural-level models of symbolic processing are highly desirable: as an important step towards the computational (re-)creation of mental capacities, as possible sources of an additional (bridging) level of explanation of cognitive phenomena of the human brain (assuming that suitably chosen ANN models correspond in a meaningful way to their biological counterparts), and also as important part of future technological developments (also see section 'The immediate vision: preparing the ground for really smart systems in the 21st century').

But while there is theoretical evidence indicating that both paradigms indeed share deep connections, how to explicitly establish and exploit these correspondences currently remains a mostly unsolved question. In the following, after a concise overview of the state of the art in the field of neural—symbolic integration in section 'Status quo in neural—symbolic integration as of 2015', as an invitation to researchers from the relevant communities two research programs are laid out which have the potential to shed light on this foundational issue: The first one, summarized in section 'Identifying and exploring differences in complexity', targets a better understanding of the empirical differences and commonalities between formalisms from the symbolic and the subsymbolic paradigm on the level of computational complexity in more scenario-specific and fine-grained ways than previously achieved. The second one, outlined in section 'Anchoring knowledge in interaction in a framework and architecture of computational cognition', gives a conceptual sketch of a research effort developing a new approach and computational architecture for the cognitively-inspired anchoring of an agent's learning, knowledge formation, and higher reasoning abilities in real-world interactions through a closed neural—symbolic acting/sensing—processing—reasoning cycle. If implemented successfully, the second program will lay the foundations for a new generation of intelligent agent systems, also giving evidence of the capacities of fully integrated neural—symbolic learning and reasoning on system level. Thus, as explained in section 'Integrating both programs: why the whole is more than the sum of the parts', when taken together both programs–besides significantly advancing the field of neural—symbolic integration–promise to greatly contribute to all four pillars and the respectively associated main scientific views of BICA identified by Stocco, Lebiere, and Samsonovich (2010). Additionally, major impact of the research programs (and the corresponding form of neural—symbolic integration) can also be expected on an immediate

---

[1] Based on results as, for instance, the ones presented by Olden and Jackson (2002), it has been argued that the inner mechanics of artificial neural networks (ANNs) can be made accessible using randomization methods and similar. While this is true when seeing ANNs as quantitative tools or means of statistical modelling, from the quite different perspective of mechanistic or explanatory knowledge about principles, rules, and processes within ANNs as part of cognitive architectures the black box character remains (with rule extraction methods, as, e.g., proposed by Andrews, Diederich, & Tickle (1995), d'Avila Garcez, Broda, & Gabbay (2001), or Zhou, Jiang, & Chen (2003), mitigating the problem only to a minimal degree).