

The Roles of Symbols in Neural-based AI: They are Not What You Think!

Daniel L. Silver¹, Tom M. Mitchell²

¹Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada B4P2R6

²Machine Learning Department, Carnegie Mellon University, Pittsburg, PA, USA, 15213

Abstract

We present a novel neuro-symbolic hypothesis and an architecture for intelligent agents that combines subsymbolic representations for symbols and concepts for learning and reasoning. We argue that symbols will remain critical to the future of intelligent systems NOT because they are the fundamental building blocks of thought, but because they characterize the subsymbolic processes that constitute thought.

In [1] we begin by defining terminology for discussing the neural encoding of symbols and concepts, and describing the key questions we seek to answer about neuro-symbolic systems. We then present relevant research results from neuroscience, behavioral (cognitive) science, and artificial intelligence, that yield evidence about the combination of symbolic and subsymbolic processing in humans and current artificial neural networks. Guided by this evidence, we present a novel neuro-symbolic hypothesis and an associated architecture meant to provide a plausible answer to the question of how humans might implement neuro-symbolic reasoning, and how future intelligent agents might be designed to do so as well.

Definitions: A **concept** is an object, a collection of objects, or an abstract idea that can be learned and represented by an intelligent agent. Concepts may range from specific physical objects (“baseball”), to a category of objects (“birds”), to very abstract and semantically complex ideas (“justice”). More complex concepts can be built out of multiple more primitive concepts (“girl riding a bike”). A **symbol** is a mark, sign or an object that represents, or refers to, some concept. For example, a logo is a symbol for a company, a word such as “peach” is a symbol for class of food, a statue of a famous person is a symbol that refers to a specific human being.


We define an agent’s internal neural activity that encodes a symbol its **symrep** for that symbol (short for the symbol’s neural representation). For example, when a person sees the word “peach”, their visual cortex generates neural activity that represents this symbol. We define an agent’s internal neural activity that encodes the concept referred to by a symbol its **conrep** (short for that concept’s neural representation). For example, when a person views the written word “peach,” their brain generates (1) neural activity that first encodes this symbol (the symrep of the letter string “peach”), and (2) neural activity that encodes the concept to which this word refers (its conrep). In contrast, when a person sees an actual peach, their brain generates neural activity that more directly encodes the concept for peach (its conrep). Depending on the context, their brain may also generate the activity for the symrep of the letter

NeSy 2023, 17th Intern. Workshop on Neural-Symbolic Learning and Reasoning, Certosa di Pontignano, Siena, Italy

✉ danny.silver@acadiau.ca (D. L. Silver); tom.mitchell@cs.cmu.edu (T. M. Mitchell)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

string “peach”. In the paper, we present results from brain imaging studies that reveal some of the properties and timing of these symrep and conrep patterns of neural activity.

A Neuro-Symbolic Hypothesis: *Symbols are critical to intelligence NOT because they are the building blocks of thought, but because they are characterizations of thought that (1) allow us to explain our subsymbolic thinking to ourselves and others and (2) act as constraints on inference and learning about the world. Symbols explain our thinking and aid our thinking, but are not the foundation of our thinking.*

We propose a neuro-symbolic architecture for an intelligent agent such as a human. It borrows from the basic diagram of an Intelligent Agent as well as work in neuroscience, cognitive psychology (specifically the ideas of *Thinking Fast and Slow* by D. Kahneman), and deep neural networks. It is also inspired by recent neuro-symbolic literature (particularly by L. Lamb and A. Garcez). The four major components of the architecture are as follows: **Sensory and Motor Subsystems** that receive raw external percepts (images, sounds, touch) from the real world and provide the appropriate percept signals to the higher order System 1 and 2 components. A **System 1 attractor network** that, given a percept signal, a symrep vector which contains recent context and attention information, and a goal-driven attention vector, learns to relax into a *conrep* activation state. This conrep will represent one or more of the previously learned concepts, each at some appropriate level of abstraction. It is semantically organized and largely grounded in sensory/motor representations with the properties: (1) two concepts with similar meanings have similar representations, and (2) operations over pairs of conrep vectors (e.g. superposition) can be performed using mapping functions. A **System 2 attractor network** that, given a percept signal, a conrep vector which contains recent context and attention information, and a goal-driven attention vector, learns to relax into a desired *symrep* activation state. The symrep will represent one or more of the previously learned symbols or some proto-symbol (discussed in the paper). System 2 is organized so that (1) symbols with similar appearance have proximal symreps, and (2) symrep operations and composition can be done with vector mapping functions (e.g. $7 \times 9 \Rightarrow 63$, “fish” + “ing” = “fishing”). A **Performance Goal Subsystem** that influences the formation of subsequent conrep and symrep given the agent’s recent perceptual input and internal state. Note that performance goals (e.g. to eat) can be closely associated with internal senses (e.g. hunger). These goals, context and attention vectors driven by recent conrep, have influence on the training and relaxation of the Systems 1 and 2 attractor networks. To some extent, agents of this architecture see what they want to see from the percepts because it helps them make sense of their world.

Conclusion: We conjecture that internal agent “self-communication” using the symrep meant for agent-to-agent communication, has become key to human intelligence because: (1) it provides a second, more abstract level of representation and reasoning which can occur in parallel with subsymbolic reasoning, and (2) it places an additional constraint on learning where prior learning act as an inductive bias for learning new symbols and concepts. Shared symbols allow us to explain and justify, internally as well as externally, our decisions and actions. And what we learn is shaped and constrained by the “lexicon” of what we recognize as symbols.

[1] D. L. Silver, T. M. Mitchell, The roles of symbols in neural-based AI: They are not what you think!, In: P. Hitzler, M. K. Sarker, A. Eberhart (Eds.), *Compendium of Neuro-Symbolic Artificial Intelligence*, IOS Press, Amsterdam, 2023. See: arxiv.org/abs/2304.13626.