Toward Human-Level Models of Minds

Philip C. Jackson, Jr.
TalaMind LLC
www.talamind.com
dr.phil.jackson@talamind.com

Abstract

A comparison of Laird, Lebiere, and Rosenbloom's Standard Model of the Mind with the 'TalaMind' approach suggests some implications for computational structure and function of human-like minds, which may contribute to a community consensus about architectures of the mind.

1. Overview

The original source for the Standard Model was a 2013 AAAI Symposium with broad representation, reflecting decades of published research. What was produced at the Symposium was followed up by focusing on three cognitive architecture systems (ACT-R, Sigma, and Soar) designed to support real-world applications. The authors of the Standard Model seek "to begin the process of engaging the international research community in developing what can be called a standard model of the mind, where the mind we have in mind here is human-like."

The TalaMind approach (Jackson 2014) is based on a review of previous research which leads to exploration of a research approach for achieving human-level AI. The thesis discusses a prototype demonstration system which is far from being ready for real-world applications, yet illustrates the potential of the research approach to achieve human-level AI. The TalaMind thesis discusses some features relevant to human-level AI which involve topics for discussion in further developing the Standard Model.

2. What is a Mind?

The paper presenting the Standard Model suggests a cognitive architecture can be equated with a hypothesis about the fixed structure of the mind. By presenting a standard model for cognitive architectures it therefore gives a standard model for the mind. This is consistent with accepting Newell & Simon's (1976) hypothesis that "A physical symbol system has the necessary and sufficient means for general intelligent action." Both the Standard Model and TalaMind include the computational capabilities of physical symbol systems, yet both also allow non-symbolic processing.

However, the Standard Model does not yet directly include some features people normally ascribe to their minds. These features involve topics for discussion in further developing the Standard Model.

3. Introduction to TalaMind

The TalaMind thesis (Jackson 2014) presents a research approach toward human-level artificial intelligence. This involves developing an AI system using a language of thought (called Tala) based on the unconstrained syntax of a natural language; designing this system as a collection of 'executable concepts' that can create and modify concepts, expressed in the language of thought, to behave intelligently in an environment; and using methods from cognitive linguistics such as mental spaces and conceptual blends for multiple levels of representation and computation. Proposing a design inspection alternative to the Turing Test, the thesis discusses 'higher-level mentalities' of human intelligence, which include natural language understanding, higher-level learning, meta-cognition and multi-level reasoning, imagination, and consciousness.

'Higher-level learning' refers collectively to forms of learning required for human-level intelligence such as learning by creating explanations and testing predictions about new domains based on analogies and metaphors with previously known domains, reasoning about ways to debug and improve behaviors and methods, learning and invention of natural languages and language games, learning or inventing new representations, and in general, self-development of new ways of thinking. The phrase 'higher-level learning' is used to distinguish these from lower-level forms of learning investigated in previous research on machine learning.

'Multi-level reasoning' refers collectively to the reasoning capabilities of human-level intelligence, including meta-reasoning, analogical reasoning, causal and purposive reasoning, abduction, induction, and deduction.

To provide a context for analysis of its approach the thesis discusses an architecture called TalaMind for design of AI systems, adapted from Gärdenfors' (1995) paper on inductive inference (see Appendix I). The TalaMind architecture has three levels, called the linguistic, archetype, and associative levels. At the linguistic level, the architecture includes the Tala language, a conceptual framework for managing concepts expressed in Tala, and conceptual processes that operate on concepts in the conceptual framework to produce intelligent behaviors and new concepts. The archetype level is where cognitive concept structures are represented using methods such as conceptual spaces, image schemas, radial categories, etc. The associative level would typically interface with a real-world environment and supports connectionism, Bayesian processing, etc. In general, the thesis is agnostic about research choices at the archetype and associative levels.

For concision, the term 'Tala agent' refers to a system with a TalaMind architecture. The architecture is open at the three conceptual levels, e.g. permitting predicate calculus, conceptual graphs, and other symbolisms in addition to the Tala language at the linguistic level, and permitting integration across the three levels, e.g. potential use of deep neural nets at the linguistic and archetype levels.

The Tala language responds to McCarthy's 1955 proposal for a formal language that corresponds to English. It enables a Tala agent to formulate statements about its progress in solving problems. Tala can represent unconstrained, complex English sentences, involving self-reference, conjecture, and higher-level concepts, with underspecification and semantic annotation. Short English expressions have short correspondents in Tala, a property McCarthy sought for a formal language in 1955.

The theoretical basis for Tala is discussed in Chapter 3 of the TalaMind thesis. Section 3.3 argues it is theoretically possible to use the syntax of a natural language to represent meaning in a conceptual language and to reason directly with natural language syntax, at the linguistic level of the TalaMind architecture. Chapter 4 discusses theoretical objections, including McCarthy's arguments in 2008 that a language of thought should be based on mathematical logic instead of natural language.

Chapter 3's analysis shows the TalaMind approach can address theoretical questions not easily addressed by more conventional approaches. For instance, it supports reasoning in mathematical contexts, but also supports reasoning about people who have self-contradictory beliefs. Tala provides a language for reasoning with underspecification and for reasoning with sentences that have meaning yet which also have nonsensical interpretations. Tala sentences can declaratively describe recursive mutual knowledge. Tala facilitates representation and conceptual processing for higher-level mentalities, such as learning by analogical, causal and purposive reasoning, learning by self-programming, and imagination via conceptual blends.

4. Discussion Topics for the Standard Model

The TalaMind thesis discusses four features relevant to human-level AI which involve topics for discussion in further developing the Standard Model.

4.1 Artificial Consciousness

The TalaMind thesis accepts the objection by some AI skeptics that a system which is not aware of what it is doing, and does not have some awareness of itself, cannot be considered to have human-level intelligence. The perspective of the thesis is that it is both necessary and possible for a system to demonstrate at least some aspects of consciousness, to achieve human-level AI. However, the thesis does not claim AI systems will achieve the subjective experience humans have of consciousness.

The thesis adapts the "axioms of being conscious" proposed by Aleksander and Morton (2007) for research on artificial consciousness. To claim a system achieves artificial consciousness it should demonstrate:

Observation of an external environment.

Observation of itself in relation to the external environment.

Observation of internal thoughts.

Observation of time: of the present, the past, and potential futures.

Observation of hypothetical or imaginative thoughts.

Reflective observation: Observation of having observations.

To observe these things, a TalaMind system should support representations of them, and support processing such representations. The TalaMind prototype illustrates how a TalaMind architecture could support artificial consciousness.

Artificial consciousness is permitted though not directly addressed in the Standard Model, since a reflective architecture is not part of the model. This was a topic over which a consensus was not reached at this point, and therefore it was omitted from the model. (Rosenbloom 2017)

Some form of artificial consciousness may be required for a consensus by scientists that an artificial intelligence has a human-level mind. People ascribe consciousness to their minds, and may expect it in artificial minds. So, this is a topic for future discussion in developing the Standard Model.

4.2 Society of Mind

The TalaMind hypotheses do not require a society of mind architecture, but it is consistent with the hypotheses and natural to implement a society of mind at the linguistic level of a TalaMind architecture. In the TalaMind prototype, a Tala agent has a society of mind in which subagents communicate by exchanging Tala concepts. Thus the TalaMind prototype simulates self-talk (mental discourse) within a Tala agent. Self-talk is an important feature people normally ascribe to their own minds.

The Standard Model includes massive parallelism within and across its modules. Whether it can support a society of mind depends on whether the parallelism included within procedural memory is adequate. However, there is not yet a consensus that society of mind is a useful paradigm for constructing general AI systems. (Rosenbloom 2017)

So, this is a topic for future consideration in developing the Standard Model.

4.3 Nested Conceptual Simulation

To support reasoning about potential future events, and counterfactual reasoning about past and present events, a Tala agent's conceptual framework should support creation and conceptual processing of hypothetical scenarios of events. A hypothetical context may include models of other agent's beliefs and goals, to support simulating what they may think and do. The TalaMind thesis uses the term 'nested conceptual simulation' to refer to an agent's conceptual processing of hypothetical scenarios, with possible branching of scenarios based on alternative events, such as choices of simulated agents.

In the TalaMind prototype, the 'farmer's dilemma' simulation shows conceptual processes in which two Tala agents (Leo and Ben) imagine what will happen in hypothetical situations, using nested conceptual simulation. Leo imagines what Ben may think and do, and vice versa. This amounts to a Theory of Mind capability within a TalaMind architecture, i.e. the ability of a Tala agent to consider itself and other Tala agents as having minds with beliefs, desires, different possible choices, etc.

Theory of Mind may require additional architectural mechanisms in the Standard Model, but that is not clear at this point. The three reference architectures for the Standard Model may support Theory of Mind without specific mechanisms for it, but that is also a topic for further discussion. (Rosenbloom 2017)

4.4 Self-Programming

People often think about how to change and improve processes. Hence a conceptual language for a system with human-level AI must be able to represent concepts that describe how to modify processes. In the TalaMind approach executable concepts can describe how to modify executable concepts. The TalaMind demonstration system illustrates this in a story simulation where a Tala agent reasons about how to change its process for making bread, the process being represented by an executable concept. This indicates the groundwork for self-programming, an important form of higher-level learning necessary for human-level AI.

The Standard Model specifies procedural memory has no direct access to itself. This prevents procedures from directly modifying procedures. However, this may not rule out self-programming: Soar and ACT-R may have demonstrated self-programming by reasoning about declarative representations of procedures in working memory and then creating corresponding procedures in procedural memory. (Rosenbloom 2017)

So, this is a topic for further discussion in developing the Standard Model.

5. Limitations of TalaMind

Of course, the TalaMind thesis does not claim to actually achieve human-level AI, or even to identify all the higher-level mentalities necessary for human-level AI. It only makes a start in this direction, and identifies many areas for future research to develop the approach. These include areas previously studied by others which were outside the scope of the thesis, such as ontology, common sense knowledge, spatial reasoning and visualization, etc. Thesis section 7.8 presents arguments in favor of the TalaMind approach over other approaches for achieving human-level AI. Involvement of the AI research community is needed for the TalaMind approach to succeed.

6. Conclusion

The TalaMind approach may be considered as a direction within the Standard Model toward Newell's vision of "a science of man adequate in power and commensurate with his complexity", and toward the vision of McCarthy, Minsky, Rochester, and Shannon who conjectured "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

What Turing wrote in 1950 is still true, "We can only see a short distance ahead, but we can see plenty there that needs to be done." Yet we have travelled far over six decades, and can now envision architectures for human-level artificial minds.

Acknowledgements

I thank Paul Rosenbloom for information referenced in Section 4 of this paper, John Laird for requesting the discussion in Appendix II, and an anonymous reviewer for comments motivating Appendix I.

Appendices

I. TalaMind's Relation to Gärdenfors (1995)

Gärdenfors (1995) discussed three ways of characterizing or describing observations, which he called the linguistic, conceptual, and subconceptual levels of inductive inference.

It is most accurate to say the TalaMind approach adapts (rather than adopts) Gärdenfors' levels by considering all of them to be conceptual levels, where concepts may be represented in different ways:

1) Linguistically
2) As cognitive categories (using methods such as conceptual spaces, image schemas, radial categories, etc.)
3) Associatively (e.g. via connectionism).

Hence TalaMind's three architectural levels are called the linguistic, archetype, and associative levels, to avoid saying only one level is conceptual.

Gärdenfors' insights remain relevant, even though his discussion of the linguistic level focused on descriptions using formal languages. However, (Gärdenfors 1995) did not discuss support for the TalaMind hypotheses at the linguistic level, and did not include elements of the linguistic level discussed in the TalaMind thesis, i.e. the Tala language, a conceptual framework for managing concepts expressed in Tala, and conceptual processes that operate on concepts in the conceptual framework to produce intelligent behaviors and new concepts. Thus (Gärdenfors 1995) did not discuss higher-level learning and other higher-level mentalities, nor aspects of minds discussed in the present paper.

II. Standard Model's Relation to TalaMind Levels

To help identify potential areas for development of the Standard Model (SM), the following paragraphs discuss how SM is related to TalaMind's three conceptual levels.

SM's processing involves symbolic data structures and production rules with pattern-matching in cognitive cycles. These data structures and production rules are symbolic expressions and would be at the TalaMind linguistic conceptual level. TalaMind is open to such symbolic expressions, though to achieve human-level AI the thesis (Jackson 2014) advocates a language (Tala) for conceptual expressions using natural language syntax, and generalizes production rules as executable concepts expressed in Tala. In the TalaMind prototype these are supported with cognitive cycles for pattern-matching of Tala expressions. Use of Tala and executable concepts may be an area for future development of SM.

SM says "Declarative memory is a long-term store for facts and concepts. It is structured as a persistent graph of symbolic relations, with metadata reflecting attributes such as recency and frequency of (co-)occurrence..." (Laird, Lebiere, and Rosenbloom 2017). This suggests symbolic relations are the primary mode for representing concepts in SM. It is not clear whether SM provides an archetype level that models cognitive categories using methods such as such as conceptual spaces, image schemas, radial categories, etc., or use of deep neural nets at the archetype level. Support of an archetype level may be an area for future development of SM.

SM's metadata about symbolic expressions could exist at the TalaMind linguistic level, though such metadata is not discussed in the thesis. Tala expressions can refer via pointers to other Tala expressions and represent statements about other expressions, supporting meta-statements in natural language syntax, which could be an area for future development of SM.

SM's declarative learning via acquisition of facts could occur at the TalaMind linguistic level, and also be supported via SM's perception component at lower levels, discussed below. At the linguistic level TalaMind focuses on higher-level learning needed for human-level AI, and is not limited to acquisition of facts or tuning of metadata. Examples of higher-level learning of declarative knowledge include: Learning by creating explanations and testing predictions, using causal and purposive reasoning; Learning about new domains by developing analogies and metaphors with previously known domains. These forms of learning may be involved in discovery of scientific theories and predictions. (To be clear, much work remains to implement higher-level declarative learning in a functioning TalaMind system.) Higher-level learning of declarative knowledge could be an area for future development of SM.

SM's procedural learning involves reinforcement learning and procedural composition. Reinforcement learning affects weights for selecting actions and procedural composition includes composition of rules and chunking. Since rules are symbolic expressions at the linguistic level, this suggests procedural learning in SM would occur primarily at the linguistic level though perhaps lower-level processes may be involved.

The TalaMind thesis discusses procedural learning at the linguistic level. As noted in Section 4.4 above, self-programming is an important form of higher-level learning. The TalaMind approach supports this by allowing executable concepts to create and modify executable concepts. The TalaMind demonstration prototype illustrates the potential for self-programming in a story simulation where a Tala agent discovers and improves a process for making bread. (Much work remains to implement self-programming in a functioning TalaMind system.) This form of procedural learning is a potential area for future development of SM.

SM's perception component could be an element at the TalaMind associative level, which in TalaMind would typically interface with a real-world environment. This corresponds to SM's statements that "Perception converts external signals into symbols and relations..." and "The standard model … does not embody any commitments as to the internal representation (or processing) of information within perceptual modules, although it is assumed to be predominantly non-symbolic in nature, and to include learning." (Laird, Lebiere, and Rosenbloom 2017)

Likewise, SM's motor component may also be an element at the associative level. SM's conversion of symbol structures into external actions is envisioned in TalaMind to happen at an interface into the associative level.

SM stipulates "More complex forms of learning involve combinations of the fixed set of simpler forms of learning". Table 1 in (Laird, Lebiere, and Rosenbloom 2017) indicates the fixed set is procedural learning at least via reinforcement and composition, plus declarative learning via acquisition of facts and metadata tuning. It seems clear this fixed set would not support the forms of higher-level learning envisioned in the TalaMind approach. This also indicates higher-level learning as a potential area for future development of the Standard Model.

References

Aleksander, I., and Morton, H. 2007. Depictive Architectures for Synthetic Phenomenology. In *Artificial Consciousness*, 67-81, ed. Chella, A. and Manzotti, R. Imprint Academic.

Fauconnier, G. 1994. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press.

Fauconnier, G. and Turner, M. 2002. *The Way We Think – Conceptual Blending and the Mind's Hidden Complexities.* Basic Books, New York.

Gärdenfors, P. 1995. Three levels of inductive inference*. Studies in Logic and the Foundations of Mathematics*, 134, 427-449. Elsevier.

Jackson, P. C. 2014. Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language. Ph.D. Thesis, Tilburg University, The Netherlands.

Laird, J. E., Lebiere, C. and Rosenbloom, P. S. 2017. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*, to appear.

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. In *Artificial Intelligence: Critical Concepts in Cognitive Science*, 2, 44-53, ed. Chrisley, R. and Begeer, S. 2000. Routledge Publishing.

McCarthy, J. 2008. The well-designed child. *Artificial Intelligence*, 172, 18, 2003-2014.

Newell, A. 1973. You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of this Symposium. In *Visual Information Processing*, 283-310, ed. Chase, W. G. Academic Press, New York.

Newell, A. and Simon, H. A. 1976. Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19, 3, 113–126.

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

Rosenbloom, P. S. 2017. Personal communication.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, 59, 433 - 460.