TalaMind White Paper

Toward Human-Level Knowledge Representation with a Natural Language of Thought

Philip C. Jackson, Jr. *

* Correspondence: dr.phil.jackson@talamind.com

August 11, 2020

Abstract: What is the nature of knowledge representation needed for human-level artificial intelligence? This position paper contends that to achieve human-level AI, a system architecture for human-level knowledge representation would benefit from a neuro-symbolic approach combining deep neural networks with a 'natural language of thought', and would be greatly handicapped if relying only on formal logic systems.

Keywords: knowledge representation; meta-knowledge; metacognition; natural language of thought; human-level AI; neuro-symbolic AI

Questions to Consider, Goals and Limitations

We are still far from representing the knowledge needed in systems that could achieve human-level artificial intelligence. It is appropriate to review what remains to be achieved, and to discuss how such knowledge could and should be represented in future systems. Therefore, the following numbered sections consider these questions, in order:

1. What is human-level intelligence? What would be human-level artificial intelligence?
2. What is knowledge? What is human-level knowledge?
3. What is the state of research on human-level AI and knowledge?
4. What are the major options to achieve human-level AI possessing human-level knowledge?
5. What are the major options for AI understanding natural language?
6. What research has been conducted toward an AI "natural language of thought"?
7. What are the arguments for and against an AI natural language of thought?
8. What is a 'TalaMind' architecture? How has TalaMind been demonstrated?
9. What future work is needed to develop the TalaMind approach?
10. Is there an 'existence proof' for eventual success of the TalaMind approach?

The goal of this position paper is to motivate cognitive scientists to take an approach to knowledge representation different from approaches previously taken, to eventually achieve human-level AI. It will be suggested that human-level AI requires 'human-level knowledge-representation', and it will be contended that this includes knowledge representation using a 'natural language of thought', supported by an architecture that includes neural networks.

This paper is based on the author's previous works ([26] *et seq.*) about a proposed approach toward eventually achieving human-level artificial intelligence, called the 'TalaMind' approach. Regarding limitations, it should be said at the outset that this position paper can only present reasons

why it is plausible this approach may achieve human-level AI and provide better support for human-level knowledge representation than other approaches.

1. What is human-level intelligence? What would be human-level AI?

To discuss the representation of knowledge needed in systems which could achieve human-level artificial intelligence, we should first discuss the nature of human-level intelligence, as it relates to requirements for systems which would exhibit human-level artificial intelligence (HLAI).

The issue of how to define human-level intelligence has been a challenge for AI researchers. Rather than define it, one might just expect to use a Turing Test to recognize it, if it is ever achieved. Some have suggested human intelligence may not be a coherent concept that can be analyzed, even though we can recognize it when we see it in other human beings. [38] (For a discussion of the nature of 'intelligence' in general, see [33], chapter 1.)

While a Turing Test may help recognize human-level AI if it is created, the test does not define intelligence nor indicate how to design, implement, and achieve human-level AI. Also, the Turing Test focuses on recognizing human-identical AI, indistinguishable from humans. It may be sufficient (and even important, for achieving beneficial human-level AI) to develop systems that are *human-like*, and understandable by humans, rather than human-identical. [28]

Therefore, an approach different from the Turing Test was proposed in [26]: to define human-level intelligence by identifying capabilities achieved by humans and not yet achieved by any AI system, and to inspect the internal design and operation of any proposed system to see if it can in principle support these capabilities, which I call *higher-level mentalities*:

- Generality
- Creativity and Originality
- Natural Language Understanding
- Effectiveness, Robustness, Efficiency
- Self-Development and Higher-Level Learning
- Metacognition and Multi-Level Reasoning
- Imagination
- Self-Awareness – Artificial Consciousness
- Sociality, Emotions, Values
- Visualization and Spatial-Temporal Reasoning
- Curiosity, Self-Programming, Theory of Mind

The higher-level mentalities together comprise a qualitative difference distinguishing human-level AI from current AI systems and computer systems in general. The following subsections briefly describe five of the higher-level mentalities. Further discussions are given in [26] and [32], beginning in §2.1.2. [1]

1.1 Generality

A key feature of human intelligence is that it is apparently unbounded and completely general. Human-level AI must have this same generality. In principle there should be no limits to the fields of knowledge the system could understand, at least so far as humans can determine.

Generality alone is not a good criterion for human-level AI, because one could describe any algorithm that searches a space of computationally general systems as having a theoretical potential to

[1] Throughout these pages the § notation is used to concisely refer to chapters and sections in [32]. For example, §2.1 refers to the first section in Chapter 2 there. Its first subsection is §2.1.1.

achieve human-level intelligence, though such a search might take eons. So, additional higher-level mentalities for human-level intelligence are discussed in §2.1.2.

Also, it is an open question whether human intelligence is actually unbounded and completely general. While we may be optimistic that it is, there are many limits to human understanding at present. Yet there is no proof at present that we cannot understand all the phenomena of nature. And it is an unsettled question whether human-level artificial intelligence cannot also do so. Jackson ([32], chapter 4) discusses theoretical objections to the potential for human-level AI.

1.2 Natural Language Understanding

Humans need a natural language like English to develop and share an understanding of the world in virtually all its aspects. To achieve human-like human-level AI, a system will need to be able to understand humans who use natural language, and need to be able to explain its thoughts to humans, using natural language, at least as well as humans can explain their thoughts to each other. Attempts to build systems that understand natural language have made substantial progress, but still founder on the problem of understanding natural language as well as humans do. No AI system today can understand English as well as a five-year-old child.

1.3 Self-Development and Higher-Level Learning

A variation of the requirement for originality is a requirement for 'self-development'. People not only discover new things, they develop new skills they were not taught by others, new ways of thinking, etc. A human-level AI must have this same capability.

More specifically, human-level intelligence includes the following higher-level forms of learning: 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; Learning by reflection and self-programming; Reasoning about thoughts and experience to develop new methods for thinking and acting; Reasoning about ways to improve methods for thinking and acting; Learning supported by invention of languages and representations; Learning by induction of new linguistic concepts; In general, learning by induction, abduction, analogy, causal and purposive reasoning.

I use the term higher-level learning to describe these collectively and distinguish them from lower-level forms of learning investigated in previous research on machine learning, such as the learning accomplished by training a neural network. Of course, there has been previous research on many of the topics mentioned above, such as analogies, causal reasoning, etc. The challenge remains to combine these methods in a unified system that can support learning as well as humans do.

1.4 Metacognition and Multi-Level Reasoning

Metacognition is "cognition about cognition", cognitive processes applied to cognitive processes. Since cognitive processes may in general be applied to other cognitive processes, we may consider several different forms of metacognition, for example: Reasoning about reasoning; Reasoning about learning; Learning how to learn….

Others have focused on different aspects of metacognition, such as "knowing about knowing" or "knowing about memory". Cognitive abilities could be considered in longer metacognitive combinations, e.g. "imagining how to learn about perception" – the combination could be instantiated to refer to a specific perception.

Such examples illustrate that natural language has syntax and semantics which can support describing different forms of metacognition. More importantly, a 'natural language of thought' could help an AI system perform metacognition, by enabling the expression of specific thoughts about other specific thoughts, specific thoughts about specific perceptions, etc. [42]

1.5 Self-Awareness – Artificial Consciousness

To exhibit human-level intelligence, a system must have some degree of awareness and understanding of its own existence, its situation and relation to the world, and its perceptions, thoughts and actions, both past and present, as well as potentials for the future. Without such awareness, a system is greatly handicapped in managing its interactions with the world, and in managing its thoughts. So, at least some aspects of consciousness are necessary for a system to demonstrate human-level intelligence.

Such awareness involves an AI system having 'artificial consciousness', which may be defined as performing the following observations:

- *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.*

This definition was proposed in [26] (p. 136), and adapted from the "axioms of being conscious" proposed by Aleksander and Morton [1]. They used first-person, introspective statements to describe these elements of artificial consciousness.

Artificial consciousness does not confront or claim to solve Chalmers' [5] Hard Problem of consciousness: There is no claim that having artificial consciousness means an AI system would have the human subjective experience of consciousness. ([32] §4.2.7)

2. What is knowledge? What is human-level knowledge?

Having defined human-level intelligence in terms of higher-level mentalities, how can we define human-level knowledge?

Published dictionary definitions are somewhat recursive, tending to define 'knowledge' at least in part by reference to 'knowing' or to what is 'known'.[2] Such definitions are not helpful for this paper.

This paper proposes and will use the following pragmatic definitions:

Knowledge is information that enables intelligence.

Human-level knowledge is information that enables human-level intelligence.

These define knowledge as information, a term defined scientifically in information theory. However, these definitions do not limit or specify how knowledge may be represented in information processing systems that might achieve human-level AI.

These definitions also leave open the topics of how knowledge enables intelligence, and how intelligence acquires knowledge. The word 'enables' is used because knowledge supports all the abilities of intelligence. Thus, in addition to knowledge that supports reasoning, and answering questions, knowledge can support performing actions or interpreting perceptions.

Human-level knowledge supports and is developed by the higher-level mentalities of human-level intelligence. Knowledge may be represented in multiple ways, even in the same system: A human-level AI might represent some of its knowledge in the weights of neural networks and other knowledge in symbolic data structures.

These definitions do not say that knowledge is perfect, static, or final. Rather, as new knowledge is developed, it may be more valid or more useful than previous knowledge, and previous knowledge may eventually be considered invalid or not useful. For example, human knowledge of the Solar

[2] For example, see https://www.merriam-webster.com/dictionary/knowledge

system became increasingly more valid and useful in the theories and observations advanced by Copernicus, Galileo, Kepler, Newton, and Einstein.

With the preceding definitions of human-level intelligence and human-level knowledge, we are ready to consider the other major questions listed at the start of this paper.[3]

3. What is the state of research on human-level AI and knowledge representation?

Although dramatic progress has been made in recent years with deep neural networks, we are still far from achieving human-level artificial intelligence. For example, no AI system yet understands natural language as well as an average five-year-old human child. No AI system can yet replicate the ability to learn and understand language demonstrated by an average child.

Knowledge representation in artificial intelligence has been a topic of study since expert systems were developed in the 1970's (and implicitly since research on artificial intelligence began in the 1950's). The breadth and nature of research on knowledge representation is indicated by the list of keywords for papers on knowledge representation and reasoning at the 2020 AAAI Conference on Artificial Intelligence:

> Knowledge Acquisition, Knowledge Engineering, Ontologies, Action, Change, and Causality, Argumentation, Automated Reasoning and Theorem Proving, Belief Change, Case-Based Reasoning, Common-Sense Reasoning, Computational Complexity of Reasoning, Description Logics, Diagnosis and Abductive Reasoning, Geometric, Spatial, and Temporal Reasoning, Knowledge Representation Languages, Logic Programming, Nonmonotonic Reasoning, Preferences, Qualitative Reasoning, Reasoning with Beliefs, Knowledge Representation.

Such research has provided useful results, and is continuing to make progress, yet we are still far from representing the knowledge needed in systems which could achieve human-level artificial intelligence: We are still far from creating an artificial intelligence that could support unconstrained interaction and dialog with humans, or could in principle understand all the knowledge humans have documented in Wikipedia, or could in principle understand or discover all the knowledge documented in scientific journals, or could learn and understand all the 'commonsense' knowledge that people use to interact with each other and the world.

4. What are the major options to achieve HLAI possessing human-level knowledge?

Logically, there are three major alternatives toward this goal:

- Purely symbolic approaches to artificial general intelligence / HLAI.
- Neural network architectures.
- Hybrid systems combining symbolic processing and neural networks.

4.1 Purely Symbolic Approaches to Artificial General Intelligence / HLAI

Based on computational universality, one could argue theoretically that purely symbolic processing is sufficient to achieve human-level AI. Over the decades, researchers have proposed a wide variety of symbolic processing approaches toward the eventual goal of achieving human-level artificial intelligence.

The term 'artificial general intelligence' (AGI) has been widely adopted for research on human-level AI. [17] Some AGI research has focused on generality, without specifically addressing other higher-level mentalities needed for human-level AI that are discussed above, e.g. natural

[3] A previous version of this paper defined knowledge as valid, useful answers to questions. That definition was objectively criticized by reviewers, leading the author to define knowledge as information that enables intelligence. The new definition subsumes information giving valid, useful answers to questions.

language, higher-level learning, metacognition, curiosity, imagination, artificial consciousness, etc. Although AGI is not limited to symbolic processing, one could discuss an approach to AGI just considering symbolic processing and symbolic representations. Although one can argue that considering generality alone is sufficient, it is reasonable to conjecture that considering generality plus other higher-level mentalities could accelerate achieving AGI and help achieve human-like AGI.

4.2 Neural Network Architectures to Achieve Human-Level AI

Based on computational generality, one can argue theoretically that neural networks are sufficient to achieve human-level AI. The technology is being applied to a wide variety of tasks in robotics, vision, speech, and linguistics. The technology is essentially domain independent.

Research on neural networks can be developed in several ways, e.g. recurrent networks and Bayesian networks, or research into models of biological neurons or topologies of neural networks similar to those in the human brain [24]. Clearly, neural networks will be an important focus of research for AI in the 21st century. There does not appear to be any theoretical reason in principle that prevents research on the wide variety of possible neural network architectures from eventually achieving a fully general human-level AI, with human-level knowledge.

However, achieving a general human-level AI via such approaches will not be easy: Human neurons are much more complex than the artificial neurons considered in conventional neural network algorithms. The human brain has about 90 billion neurons, and about 100 trillion connections (synapses) between neurons. It may not be feasible to adequately simulate real neurons in such orders of magnitude by a computer system, perhaps even in this century, although research projects have been undertaken in this direction, e.g. [43].

Also, the development of human intelligence within the brain of a child follows a different path from the training sequence of a conventional neural network, leveraging natural language communication and interaction with other humans.

Finally, if human-level AI is achieved solely by relying on neural networks then it may not be very explainable to humans: Immense neural networks may effectively be a black box, much as our own brains are largely black boxes to us. It will be important for a human-level AI to be more open to inspection and more explainable than a black box. These factors suggest that research on neural networks to achieve human-level AI should be pursued in conjunction with other approaches that support explanations in a natural language like English, support a child-like learning process, and avoid complete dependence on neural nets by allowing an AI system to use other computational methods when neural nets aren't needed.

4.3 Hybrid Architectures Combining Symbolic Processing and Neural Networks

There is no reason in principle why hybrid architectures cannot be developed, combining symbolic processing and neural networks to support eventually achieving human-level AI. Indeed, there could be substantial advantages in developing hybrid architectures: Each approach could augment the other. Symbolic processing could support representing, reasoning, and learning with sentential structures, networks, contexts, etc. Neural networks could support learning, representing, and recognizing complex patterns and behaviors that are not easily defined by symbolic expressions. (This is not a new idea, cf. [19].) This paper will advocate a class of hybrid architectures called the 'TalaMind architecture'. [32] The approach advocated here belongs in the camp of "neuro-symbolic AI" research, though it has not yet been generally accepted by that camp. [7] [21]

5. What are the major options for AI understanding natural language?

Understanding natural language was listed in section 1 above as one of the higher-level mentalities of human-level intelligence. In many ways it is a key higher-level mentality because it supports other higher-level mentalities.

The major options for AI systems to understand natural language are parallel to the options for AI systems to achieve human-level intelligence: AI systems could use purely symbolic programming methods for processing and understanding natural language, or rely entirely on neural networks, or use hybrid approaches combining symbolic processing and neural networks.

If we focus just on the symbolic processing methods, there are two major alternatives to discuss.

5.1 Treating Natural Language as External Data for an AI System

The first alternative for symbolic processing of natural language is to treat natural language expressions as external data, and to use other, internal symbolic languages for representing thoughts and for specifying how to process, interpret and generate external natural language expressions. It has been a traditional approach to translate natural language expressions into a formal language such as predicate calculus, frame-based languages, conceptual graphs, relational tuples, etc., and then to perform reasoning and other forms of cognitive processing, such as learning, with expressions in the formal language. Some approaches have constrained and "controlled" natural language, so that it may be more easily translated into simpler formal languages, database queries, etc.

5.2 Using Natural Language as a Language of Thought in an AI System

The second major alternative is to represent natural language expressions as internal data structures and to use natural language itself as an internal symbolic language for representing thoughts, and for describing (at least at a high level) how to process thoughts, and for interpreting and generating external natural language expressions.

This approach is what I describe as implementing a '*natural language of thought'* in an AI system. It will be further discussed throughout this paper. Section 8 discusses implementation and demonstration in a prototype system. Additional discussions are given in [26] *et seq*.

Other symbolic languages could be used internally to support this internal use of natural language, e.g. to support pattern-matching of internal natural language data structures, or to support interpretation of natural language data structures. This approach could also be combined with neural networks, in hybrid approaches for processing natural language. Yet in this approach the data structures representing natural language expressions are the general high-level representations of thoughts. For domains like mathematics, physics, chemistry, etc. an AI system might use additional symbolic languages to help represent domain-specific thoughts.

This approach involves more than just representing and using the syntax of natural language expressions to represent thoughts: It also involves representing and using the semantics of natural language words and expressions, to represent thoughts. [32, 34] This is further discussed in section 8 below.

There is not a consensus based on analysis and discussion among scientists that an AI system cannot use a natural language like English as an internal language for representation and processing of thoughts. Rather, in general it has been an assumption by AI scientists over the decades that computers should use formal logic languages (or simpler symbolic languages) for internal representation and processing within AI systems. Yet it does not appear there is any valid theoretical reason why the syntax and semantics of a natural language like English cannot be used directly by an AI system for its language of thought, without translation into formal languages, to help achieve human-level AI ([32], pp. 156-177).

6. What research has been conducted toward an AI natural language of thought?

Historically, it appears there have been very few research endeavors directly toward developing an AI natural language of thought, though there have been endeavors in related directions.

6.1 Research Directly Toward an AI Natural Language of Thought

For his research in the Dartmouth Summer Research Project on Artificial Intelligence, McCarthy [44] proposed to create a computer language that would have properties similar to English. The artificial language would allow a computer to solve problems by making conjectures and referring to itself. Concise English sentences would have equivalent, concise sentences in the formal language. McCarthy's envisioned artificial language would support statements about physical events and objects and enable programming computers to learn how to perform tasks and play games.

Although McCarthy proposed in 1955 to develop a formal language with properties similar to English, it is not clear what research he conducted in this direction at that time. Beginning in 1959 his papers concentrated on use of predicate calculus for representation and inference in AI systems, while discussing philosophical issues involving language and intelligence. In 2008 he published a paper arguing that a natural language like English would not work as a language of thought for a human-level artificial intelligence. His paper is discussed briefly in section 7.2.1 below, and in more detail by [32] §4.2.5.

In 1979, Noah Hart wrote a senior thesis[4] on the use of natural language syntax to support inference in an AI system. He described a Lisp program that used list structures to represent syntax and semantics of English sentences, which could answer questions about knowledge represented in such list structures.

In 2014, I received a doctorate for a thesis [26] that advocated developing an AI system using an internal language (called Tala) based on the unconstrained syntax of a natural language (English), and taking a principled approach toward supporting the unconstrained semantics of natural language. A prototype demonstration system was developed using Tala as a symbolic language for representing information and procedures. In Tala, natural language expressions are symbolic data structures (represented as hierarchical list structures) that can represent natural language syntax and semantics.[5] The prototype's cognitive cycle used pattern-matching of Tala expressions for information and procedures. The thesis proposed a hybrid architecture called TalaMind for systems to achieve human-level AI, which when fully developed would include an associative level for neural networks. For concision, a system with a TalaMind architecture is called a 'Tala agent'. The following pages will provide additional discussion of TalaMind, as appropriate to addressing specific topics of this paper. Section 8 below discusses details of the TalaMind architecture and prototype system. [32]

6.2 Research Related to an AI Natural Language of Thought

6.2.1 Research on Languages of Thought by Afzal Ballim and Yorick Wilks

A discussion of languages of thought was provided by Ballim and Wilks (1991). The discussion appears consistent with each author's writings to the present. They reason in favor of humans having and using languages of thought, and in favor of AI systems using such languages to emulate human thinking. They note that natural languages are essential for "keeping the behaviors, beliefs, goals, etc. of different organisms in harmony to some acceptable degree" and that a natural language is "the only symbolic evidence we have on which to base our construction of the beliefs and attitudes" of humans.

They also write:

> "Our view (Wilks, 1982) is that there is no 'escape theory' for natural languages to another nonsymbolic domain of entities that somehow 'justify' meaning ... Meaning remains to be found in other symbols...a position that owes much to the views of Wittgenstein and Quine... To say that meanings are other words or symbols is, of course, not enough...among those symbols are essentially symbols for beliefs, goals, and plans."

---

[4] Hart's senior thesis was unpublished and is unavailable.

[5] Although much semantics can be represented linguistically via natural language expressions, a Tala agent's full representation of semantics requires representations at the archetype and associative levels of a TalaMind architecture. This is discussed further in section 8.1 of this paper and in [32] and [34].

Ballim and Wilks also observed that there is no way to show that all humans use the same language of thought, and no way to show that humans use natural languages as languages of thought. Hence, they focused on developing symbolic languages and formal logics that AI systems could use to represent beliefs, goals, and plans.

However, their arguments do not preclude taking the TalaMind approach, to develop an AI system that supports representing beliefs, goals, and plans using an internal natural language of thought within a conceptual framework. Nor do they preclude an AI system's natural language of thought supporting multiple human natural languages, si te das cuenta de lo que quiero decir. [35]

The TalaMind approach also takes a different view of the argument that "meanings are other words or symbols": While many meanings may be given by other words or symbols, or by other natural language expressions, for TalaMind many meanings may also be in the associative layer, and correspond to patterns recognized by neural networks.

6.2.2 Research on 'Innerese' by Patrick Winston and Colleagues

Research in the Genesis group at MIT by Patrick Winston and his colleagues studied how an AI system could use "a universal symbolic language (called Innerese) that can represent any problem and goal". (Thrush and Winston [58])

Innerese was described as having the properties of a language of thought. Using the START natural language system [41], English expressions are translated to Innerese expressions, which are hierarchically nested `<subject relation object>` triples. In principle, Innerese expressions like `<wrench on-top-of table>` could be generated from a robot's perceptions of the world, i.e. without being translated from English.

From the perspective of this position paper, the research on Innerese illustrates the potential value for AI systems having languages of thought, and does not preclude a potential value for human-level AI systems having a natural language of thought based on English or other human natural languages. Likewise, the TalaMind approach to reasoning directly with the syntactic structure of English expressions does not preclude reasoning with simpler languages like Innerese, when useful. Further comparisons of Innerese and Genesis with Tala and the TalaMind approach would be topics for a future paper.

6.2.3 Natural Logic Research

Natural Logic research has studied how natural language syntax can be analyzed to support logically valid reasoning. Kartunnen [40] gave an overview of research on Natural Logic and summarized the history of research as described by van Benthem [3], tracing the topic from Aristotle, to William of Ockham, to Charles Sanders Peirce, to modern times.

Natural Logic research has shown that certain kinds of reasoning can be performed directly without translation of natural language syntax into predicate calculus or a higher-order logic. To this extent, Natural Logic research supports the approach advocated by the TalaMind approach and this position paper. The TalaMind approach could be considered as a kind of Natural Logic since it involves reasoning directly with natural language syntax.

Regarding limitations of research on Natural Logic, Kartunnen [38] wrote that:

> "The shortcomings and unsolved problems with Natural Logic … are challenging in any framework for semantics. The common thread of the inference problems discussed below is the need to take into account pragmatic factors, the context of use, and even the perceived intent of the speaker. … people make inferences that go beyond what the sentence logically entails or presupposes."

The TalaMind approach differs from previous research on Natural Logic by taking a different approach to the problems Kartunnen identified, supporting consideration of pragmatics. [34, 35]

For example, Kartunnen [38] discussed interpretations of the sentence *Paul was not smart to take the middle piece*. Some people interpret the sentence as implying that Paul was not smart and took the middle piece. Others interpret it as implying that Paul was not smart and did not take the middle piece.

There is no single 'right' interpretation for the sentence. People can be influenced to take different interpretations by using different variations of the sentence, but Kartunnen concluded "some very basic inferences such as whether the event described by an infinitival complement happened or not depend on opinions that are not part of the literal meaning of the sentence. This is a difficult problem for compositional semantics and for Natural Logic as well."

It is not a fundamental problem for the TalaMind approach, because in principle a Tala agent could consider both interpretations of the sentence and not assume that either interpretation is true. A Tala agent could use contextual knowledge to identify the intended interpretation, or ask for clarification. In principle, a Tala agent could use relevance theory [61] to support pragmatic disambiguation. Relevance theory is discussed further in section 7.2.4 below, and in [35].

Kartunnen [39] also gave several examples of different interpretations for sentences using the phrase "will be lucky":

Wong Kwan will be lucky to break even.
Wong Kwan was lucky to break even.
Wherever she ends up, they will be lucky to have her.
In fact you will be lucky to see any traffic at all.
In fact you will be lucky to see at least some traffic.

Such examples illustrate that "will be lucky to X" sometimes means X is likely and sometimes means X is unlikely, depending on other information in the sentence and in the phrase X, and depending on context information, and ultimately on commonsense.

In the TalaMind approach, a Tala agent would not be constrained to assume that a linguistic phrase can only have a single, standard interpretation in all usages. In principle, a Tala agent would be able to understand what is intended by a usage of the phrase "will be lucky to X", using the same kinds of information that humans use to understand it. (Viz. [32] §3.6.6.2, §3.6.3, and [35].)

The TalaMind approach may provide the "surfacy" Natural Logic sought by van Benthem [3]. Yet I should be clear in saying that much work remains to develop TalaMind systems which achieve these possibilities. This position paper can only present theoretical arguments motivating future research in this direction. Some additional discussion of research on Natural Logic is given in [32] §2.3.7.

7. What are the arguments for and against an AI natural language of thought?

7.1 Arguments For an AI Natural Language of Thought

7.1.1 Moravcsik's Arguments for Natural Language

Moravcsik [50] gave reasons why natural language is not a formal logic language. Moravcsik's arguments are in effect, arguments in favor of AI systems using a natural language of thought, rather than formal logic. His arguments imply obstacles for the translation of natural language to and from formal logic, and for representation in formal logic of human-level knowledge that is normally expressed in natural language. They imply obstacles for using formal logic within AI systems to achieve human-level understanding of natural language. (The issues discussed above that were identified by Kartunnen for Natural Logic are also, in effect, reasons supporting an AI natural language of thought.)

Moravcsik considered universal quantification and the meaning of the word "all" in sentences of the form "all A's are B's", asking whether sentences of this form imply that A's exist. He discussed three different interpretations of universal quantification: 1) If "all A's are B's" is true, then at least one A exists;[6] 2) The sentence "all A's are B's" presumes that A's exist, and the sentence does not have a truth value if no A's exist;[7] 3) The sentence "all A's are B's" may be true even if no A's exist.

[6] Moravcsik notes that Aristotle advocated this on the grounds that sciences are about things that exist.
[7] Moravcsik cites Strawson [57] for this viewpoint.

As an example of the first interpretation, Moravcsik observed that if a door-to-door salesman says "all the other children in this neighborhood have this toy" then the implication is that there are other children in the neighborhood, and the salesman may only be considered truthful if there are other children in the neighborhood and they all have the toy.

As an example of the second interpretation, Strawson ([57], p. 344) wrote that "A literal-minded and childless man asked whether all his children are asleep will certainly not answer 'Yes' on the ground that he has none ; but nor will he answer 'No' on this ground." Thus, Strawson said the sentence "All my children are asleep" does not have a truth value, for the childless man.

In the third 'lawlike' interpretation, "all A's are B's" may be considered true, even if no A's exist. This interpretation is useful for reasoning about things independently of whether they exist. Moravcsik noted that a scientific statement "All pure water is H_2O" may be considered true, even if there does not happen to be any perfectly pure water, and in the day-to-day world if a law says "all trespassers will be prosecuted", then the statement may be considered true even if there are no trespassers. Moravcsik referenced Goodman's discussion of lawlike statements. [18]

Thus, Moravcsik suggested there is a basis for multiple senses of quantifiers in natural languages. This could be supported in an AI system's natural language of thought – related topics are discussed in [32] §3.6.

Although Moravcsik identified issues that handicap formal logic in representing the knowledge and reasoning that we normally express in natural language, he wrote that this "does not prevent us from specifying logical expressive power in context, and thus reasoning deductively within natural language, with contextual semantic specifications given to some structures." ([50], p. 84) Thus, he acknowledged our ability to use natural language for reasoning. In principle, it seems clear he would have supported the use of a natural language of thought by AI systems, though it does not appear he specifically discussed this idea.

7.1.2 Jackson's Analysis of Issues for a Natural Language of Thought in Human-Level AI

I allocated a chapter in [26] to discussing how a natural language of thought called Tala could support achieving human-level artificial intelligence. Quoting from the summary of Chapter 3:

> "This analysis showed that the TalaMind approach allows addressing theoretical questions that are not easily addressed by other, 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 sentences can express meta-concepts about contexts, such as statements about consistency and rules of conjecture. And the TalaMind approach 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."

Chapter 5 of [26] presented a prototype design for the syntax of the Tala conceptual language. This syntax was fairly general and flexible, and covered many of the issues discussed by Hudson [23] for Word Grammar dependency syntax.

7.1.3 Summary

Moravcsik [50] and Jackson [26, 32] gave arguments that support AI systems using a natural language of thought, and not relying solely on formal logic. The issues discussed above in section 6.2.2 that were identified by Kartunnen [40] for Natural Logic are also, in effect, reasons supporting an AI natural language of thought.

7.2 Arguments Against an AI Natural Language of Thought

7.2.1 McCarthy's 2008 Paper

In 2008, McCarthy published a paper giving arguments that natural languages like English would not work as languages of thought, both for humans and for AI systems with human-level intelligence. He said it would be "appropriate" for a robot's language of thought to be based on logic. His arguments are addressed in detail by [32] §4.2.5, finding that McCarthy was mistaken in discounting natural language as a basis for an AI system's language of thought. His arguments rely on an incorrect assumption that the characteristics of external, public spoken or written natural language would necessarily obtain for an AI system having an internal language of thought with syntax and semantics based on natural language.

7.2.2 McShane and Nirenburg's 2012 Paper

McShane and Nirenburg ([48] pp.11-13) discussed the idea of using natural language as a knowledge representation language, and rejected it as impractical, writing "engines for reasoning directly in NL do not exist and may never exist." Unaware of their paper, I gave a detailed discussion ([32], chapters 5 and 6) of a prototype demonstration system for reasoning directly in natural language. This demonstration system is described in section 8.2 below.

McShane and Nirenburg [48] wrote that Wilks had advocated using natural language as a knowledge representation language, but said "ambiguity of representation was not a central issue" for Wilks, and "the types of applications Wilks has in mind rely only partially on addressing issues of text meaning." In recent correspondence with me, Wilks disagrees with these statements in [48].

Jackson ([32], §2.2.1) notes a paper giving a dialog between Nirenburg and Wilks [52] regarding four questions, one of which is "Are representation languages (RLs) natural languages (NLs) in any respect?" In the dialog, Wilks appears to support the affirmative, and Nirenburg the negative, though with various agreements, disagreements, and uncertainties.

Although the previous paragraphs suggest differences of perspective, it may be possible to pursue research on a natural language of thought leveraging the OntoAgent framework described by [48].

7.2.3 The Argument that What We Say Is Only the Tip of the Iceberg

A reviewer of a previous version of this paper gave the following criticism:

> "The author seems to assume that since we, as people, can talk about something (e.g. reflection, ambiguity, polysemy) and that we use natural language to do it, that it only makes sense that a computer could do the same. He neglects the fact that language is used intentionally in an interpersonal context, and that what we explicitly say is only the tip of the iceberg of what we communicate. I highly recommend reading the works of Michael Tomasello to gain a larger perspective."

I do not say (or assume) that it "only makes sense" that a computer could use a natural language of thought, just because people use natural language for communication. Rather, I say that there does not appear to be any valid theoretical reason why a computer could not use a human natural language as its language of thought, in the manner described by [26] *et seq*.

I do not neglect the fact that humans use natural language intentionally in an interpersonal context, and that what we explicitly say is only the tip of the iceberg of what we communicate. These facts are challenges for AI systems to communicate with humans using natural language, but they do not theoretically preclude an AI system from using a natural language as its internal language of thought. The "tip of the iceberg" issues for communication are essentially the ambiguity issues for understanding natural language, discussed in the next section.

Tomasello's writings appear to focus on the interpersonal use and acquisition of natural language by humans. I do not claim to be an expert on his writings, but I have not found any which specifically contradict the theoretical possibility that an artificial cognitive system could in principle use a natural language as a language of thought in the manner described by [32].

7.2.4 The Argument that Ambiguity is a Theoretical Barrier for an Artificial NLoT

It may be objected that ambiguity is a theoretical barrier to the use of a natural language of thought by an artificial cognitive system.

Such an objection is itself ambiguous, if it does not specify how ambiguity is a theoretical barrier. Yet this objection has only been expressed to the author without being specific. So, this paper can only give a general response while considering specific issues as well as possible, due to page limits this paper already greatly exceeds. A more extensive discussion is given by [35].

Sowa [55] noted that ambiguity in natural language has a major benefit: Ambiguity allows us to communicate about anything without having to be specific about everything:

> "What makes formal logic hard to use is its rigidity and its limited set of operators. Natural languages are richer, more expressive, and much more flexible. That flexibility permits vagueness, which some logicians consider a serious flaw, but a precise statement on any topic is impossible until all the details are determined. As a result, formal logic can only express the final result of a lengthy process of analysis and design. Natural language, however, can express every step from the earliest hunch or tentative suggestion to the finished specification."

This advantage of ambiguity in natural language goes beyond topics such as logical analysis and design of systems: It extends to our day-to-day interactions and activities in virtually all domains.

Also, Sowa [55] noted that formal logic is just a notation based on natural language:

> "Aristotle developed formal logic as a systematized method for reasoning about the meanings expressed in ordinary language. …In the 19th and 20th centuries, mathematicians took over the development of logic in notations that diverged very far from its roots, but every operator of any version of logic is a specialization of some word or phrase in natural language… every step of every proof in any formal logic can be translated to an argument in ordinary language that is just as correct and cogent as the formal version."

And it should be noted that formal logic is also inherently ambiguous: Any formal logic theory has undefined terms, which are ambiguous outside the theory. Humans may interpret these terms as having meanings if the theories are applied to the real world, but such meanings depend on human understanding, which is generally achieved via natural language. Using natural language, we can name and describe events and things we do not know how to fully define or explain with formal theories.

Wilson and Sperber [61] discuss how pragmatic reasoning according to relevance theory can support disambiguation for many natural language usages, including rhetoric, metaphors, irony, hyperbole, etc. They write (p. 20) that relevance theory can provide a unified account of how people understand literal, vague, loose, and figurative meanings. They describe disambiguation as a search that follows a path of least effort (p. 7) to find an interpretation for an utterance that has optimal relevance to the people communicating, without having to consider multiple other interpretations. This least-effort search enables disambiguation to be achieved rapidly.

In principle, the pragmatic reasoning Sperber and Wilson discuss for human disambiguation of natural language could also be supported by an artificial cognitive system possessing a comprehensive lexicon, conceptual ontology, encyclopedic and commonsense knowledge. With these resources, the ambiguities of natural language would also not be a theoretical barrier for an AI system to use a natural language like English as its internal language of thought. [35]

Moreover, AI systems would not encounter ambiguity in a natural language of thought as pervasively as ambiguity is encountered when natural language is used for communication between humans. When an AI system perceives an object or event in its environment, it could create an internal pointer to its perception, and use the internal pointer as a referent in any thoughts it creates about its perception of the object or event. If its thoughts are represented as structures in a natural language of thought, such internal pointers to its perceptions would allow the system to avoid having ambiguities of reference in its thoughts.

The AI system could also create pointers to its thoughts if its thoughts are represented internally as structures in a natural language of thought. Such pointers could remove ambiguities from its

thoughts about its thoughts, supporting metacognition.

A natural language of thought is also the most natural, general way for expressing the encyclopedic knowledge and linguistic definitions of word senses that are needed in a human-level AI to support disambiguation of natural language utterances received from the environment. Again however, there is no requirement that a human-level AI must only use a natural language of thought for its representations: Its architecture should be open to a wide range of representation methods, such as conceptual graph structures (Sowa [55]), mathematical and scientific notations, conceptual spaces (Gärdenfors, [16]), mental models (Johnson-Laird, [37]), neural networks, etc.

Finally, there are situations where preserving ambiguities is a virtue: It may not be necessary to disambiguate everything in a sentence to achieve an optimal interpretation. For example, consider:

> In most democratic countries most politicians can fool most of the people on almost every issue most of the time. (Hobbs, [22])

A normal interpretation is that the quantifiers vary across countries, politicians, people, and issues in an unspecified manner that is not important. Yet, technically the sentence can have interpretations where quantifiers do not vary. For example, one interpretation is that people between the ages of 10 and 80 can be fooled about the same majority set of issues across a certain majority of democratic countries, but only between 2 a.m. and 5 p.m. of each day. Such an interpretation is not supported by commonsense, but it is logically permitted. A human-level AI should not need to consider it, if it does not affect how the example sentence is being discussed. With a natural language of thought, an AI system could represent and reason about a hierarchical syntax structure for the example sentence, without needing to consider such interpretations.

The bottom line of these remarks is that there is not a valid theoretical reason why ambiguity is a barrier to the use of a natural language of thought by an artificial cognitive system. Rather, there is synergy in using a natural language of thought to support disambiguation of natural language expressions received from the environment, and there are theoretical reasons why support of ambiguity is a virtue of a natural language of thought for an artificial cognitive system.

7.2.5 The Argument that Humans Do Not Use a Natural Language of Thought

It may be objected that human-level AI cannot or should not be achieved by using an artificial natural language of thought, since it is not clear that humans use a language of thought.

While the research of Fernyhough [12, 13] on self-talk suggests to me that humans might use natural languages of thought, this paper is not embracing or advocating specific arguments about languages of thought in humans. Thus, it is not endorsing Fodor's [14] arguments, nor the alternative proposed by Schneider [54] for a language of thought developed to be compatible with cognitive science and neuroscience, nor Carruthers' arguments [4] for natural language playing a role in human cognition. Such arguments are interesting, yet this paper focuses only on the nature of a language of thought for an artificial cognitive system that could arguably achieve human-level artificial intelligence.

In this regard, an argument this paper does accept is Jackendoff's reasoning [25] that some concepts must be expressed as sentences in a mental language: Since there are an effectively unlimited number of natural language sentences that humans could understand, and our brains are finite, it follows that "sentential concepts" must be represented internally within the mind as structures within a combinatorial system, or language. For sake of discussion, we may call this internal language a 'mentalese', again without embracing specific arguments by Fodor or others about the nature of mentalese in humans.

Plausibly however, the expressive capabilities of natural languages should be matched by expressive capabilities of mentalese, or else the mentalese could not be used to represent concepts expressed in natural language. The ability to express arbitrarily large sentence structures and to express sentences that refer to sentences is plausibly just as important in a mentalese as it is in English. The ability to metaphorically express ideas across arbitrary, multiple domains is plausibly just as important in a mentalese as it is in English.

This is not to say mentalese would have the same limitations as spoken English, or any particular natural language. In a natural language of thought, sentences could have graphical structures not physically permitted in speech.

7.2.6 Summary

A variety of arguments against an AI system using a natural language as its language of thought have been considered. Each of these arguments has been shown to be invalid.

8. What is a TalaMind architecture? How has TalaMind been demonstrated?

8.1 The TalaMind Approach and Architecture

The 'TalaMind' approach was proposed by this author for research toward eventually achieving human-level artificial intelligence. [32] The approach is summarized by three hypotheses:

I. Intelligent systems can be designed as 'intelligence kernels', i.e. systems of concepts that can create and modify concepts to behave intelligently within an environment.

II. The concepts of an intelligence kernel may be expressed in an open, extensible conceptual language, providing a representation of natural language semantics based very largely on the syntax of a particular natural language such as English, which serves as a language of thought for the system.

III. Methods from cognitive linguistics may be used for multiple levels of mental representation and computation. These include constructions, mental spaces, conceptual blends, and other methods. [10, 11]

The first hypothesis essentially describes the 'seed AI' approach in AGI. [61] The second hypothesis conjectures that a language of thought based on the syntax and semantics of a natural language can support an intelligence kernel achieving human-level artificial intelligence. The third hypothesis envisions that cognitive linguistics can support multiple levels of cognition.

The TalaMind architecture (Figure 1) has three levels of conceptual representation and processing, called the linguistic, archetype, and associative levels, adapted from Gärdenfors' paper [15] on levels of inductive inference. He called these the linguistic, conceptual, and associative levels, but the perspective of the TalaMind approach is that all three are conceptual levels. For example, the linguistic level includes sentential concepts. Hence the middle level is called the archetype level, to avoid implying it is the only level where concepts exist.

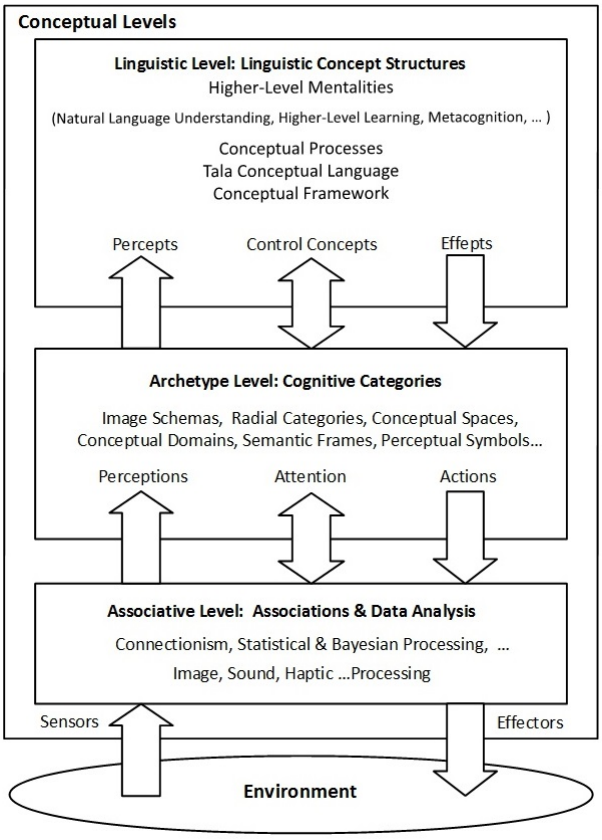

Figure 1. The TalaMind Architecture

At the linguistic level, the architecture includes a natural language of thought called Tala, 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. Conceptual processes can be implemented with 'executable concepts' expressed in Tala,

which can create and modify executable concepts. The archetype level is where cognitive categories 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 support neural networks, Bayesian processing, etc. At present, the TalaMind approach does not prescribe specific research choices at the archetype and associative levels. Words and expressions at the linguistic may level to refer to concepts at the archetype level or percepts at the associative level.

For concision, the term 'Tala agent' refers to a system with a TalaMind architecture. The architecture is open at the three conceptual levels, permitting conceptual graphs, predicate calculus, and other formal languages in addition to the Tala language at the linguistic level, and permitting integration across the three levels, e.g. potential use of deep neural networks at the linguistic and archetype levels. The TalaMind architecture is actually a broad class of architectures, open to further design choices at each level.

These topics are further discussed in [32] and related writings by the author, listed in the References.

Note that a conceptual representation may span levels and forms of representation, e.g. a linguistic concept structure may reference a cognitive concept structure. Also, some authors may disagree with this placement at different levels. Thus, Fauconnier and Turner might argue mental spaces and conceptual blends should be at the archetype level. While conditional probabilities fit the associative level, Bayesian networks may represent semantics of sentences at the linguistic level in future research. Within the scope of this paper, precisely how concepts are represented in the archetype and associative levels is not crucial. A Tala agent may not need to include all the different forms of concept representation listed above, particularly at the archetype level, since these overlap in representing concepts. Ways to unify representations within or across the three levels may be a worthwhile topic for future research.

In proposing development of a natural language of thought called Tala, based on the syntax and semantics of English, the TalaMind approach does not prescribe an approach for structuring the Tala lexicon. It could leverage network and inheritance work by previous researchers. It is open to use of a generative lexicon and uses grammatical constructions for representing and extending natural language syntax and semantics. (§3.6.3.13)

When fully developed the Tala language would support as many concepts and relations as there are word senses in a natural language like English, i.e. many thousands. Tala would enable expressing an unlimited set of phrases and sentential concepts. These might be considered as 'knowledge constructs', for comparison with conventional knowledge-representation systems. The topic of 'primitive' words for the Tala language is discussed in §3.6.8.

The TalaMind hypotheses do not require it, but it is consistent and natural to have a society of mind at the linguistic level of a TalaMind architecture. The term 'society of mind' is used in a broader sense than the approach described by Minsky [49]. This broader, generalized sense corresponds to a paper by Doyle [8], who referred to a multiagent system using a language of thought for internal communication, although Doyle did not discuss a 'natural language of thought'.

A Tala expression is a (potentially reentrant) multi-level list structure representing the dependency parse-tree (syntax) of a natural language expression. (§3.5.2) For example, the sentence "Can you turn grain into food for people?" could be represented by:

```
(turn
  (wusage verb)
  (modal can)
  (sentence-class question)
  (subj you)
  (obj (grain wusage noun)))
  (into
     (food (wusage noun)
       (for (people (wusage noun))
            ))))
```

If a natural language expression is heard or seen in the environment, then a Tala agent will have conceptual processes for constructing alternative syntactic and semantic interpretations of the serial word expression, to understand which interpretation is intended in the current context, and reason with the interpretation. These processes may result in asking for clarification, of course.

However, when Tala expressions are created and processed internally within a Tala agent, they are created and processed as syntactic structures. There is no need within a Tala agent to convert internal syntactic structures to and from linear text strings. Such internal processing also need not involve disambiguation since Tala expressions can include pointers to word senses and referents.

In saying the Tala language of thought is based on the unconstrained syntax of a natural language, it should be clear that it is not a 'controlled natural language'. Using Tala, a human-level AI should be able to represent and understand both grammatical and ungrammatical natural language, though internally it might translate ungrammatical expressions into grammatical ones. It should be able to represent and understand metaphors, metonyms, anaphora, idioms, multiple negatives, etc. These and other issues are discussed in Chapters 3 and 5 of [32].

8.2 The TalaMind Prototype Demonstration System

To illustrate further how human-level AI may eventually be achieved by developing systems that can represent human-level knowledge using a natural language of thought, this section describes the design, processing, output, limitations and evaluation of the TalaMind prototype demonstration system.

Chapter 5 of [32] presents the design for the TalaMind prototype demonstration system. It is a functional prototype in which two Tala agents, named Ben and Leo, interact in an environment that is simulated with symbolic expressions. Each Tala agent has its own prototype TalaMind conceptual framework and conceptual processes. To the human observer, a simulation is displayed as a sequence of English sentences, which is in effect a story describing interactions between Ben and Leo, their actions and percepts in the environment, and their thoughts.

For the thesis, two stories were simulated in which Ben is a cook and Leo is a farmer. The first is a story in which Ben and Leo discover how to make bread. In the second story, Ben and Leo agree to an exchange of wheat for bread and then perform the exchange, after thinking about what may happen if the agreement is fulfilled or broken.

Chapter 5 of [32] presents a design for the syntax of the Tala conceptual language which is fairly general and flexible, and covers many of the issues discussed by Hudson [23] for Word Grammar dependency syntax, although the TalaMind approach is not limited to use of dependency syntax. Such coverage is described to suggest that Tala's syntax could eventually be comprehensive for English, though developing a comprehensive Tala syntax for English is itself a very large effort that could occupy multiple researchers.

The Tala syntax also supports non-grammatical natural language expressions, because people frequently use non-grammatical language and a human-level AI needs to be able to represent and try to understand whatever people say. The Tala syntax is somewhat non-prescriptive, open and flexible, e.g. by making parts of speech optional. ([32], p. 77)

The prototype system design includes a conceptual framework and conceptual processes at the linguistic level of a TalaMind architecture. The conceptual framework includes prototype representations of perceived reality, a Tala lexicon, encyclopedic knowledge, mental spaces and conceptual blends, executable concepts, grammatical constructions, and event memory. The prototype conceptual processes include interpretation of executable concepts with pattern-matching, variable binding, conditional and iterative expressions, transmission of internal speech acts between subagents, conceptual blending, and composable interpretation of grammatical constructions. The prototype implements these features only to a preliminary extent.

The TalaMind prototype demonstration system includes pattern-matching logic for Tala expressions to support inference with natural language syntax. Executable concepts and word senses are represented in the prototype Tala language.

Table 1. Output from the 'Discovery of Bread Story Simulation' (Jackson, 2014).

Time step	Event
1...1	Leo has excess grain.
1...1	Leo thinks Leo has excess grain.
1...2	Leo tries to eat grain.
1...4	Leo asks Ben can you turn grain into fare for people?.
1...7	Ben examines grain.
1...8	Ben thinks wheat grains resemble nuts.
1...8	Ben imagines an analogy from nuts to grain focused on food for people.
1...8	Ben thinks grain perhaps is an edible seed inside an inedible shell.
1...17	Ben mashs grain.
1...20	Ben thinks dough is too gooey.
1...21	Ben bakes dough.
1...23	Ben tries to eat flat bread.
1...28	Ben thinks people would prefer eating thick, soft bread over eating flat bread.
1...29	Ben thinks how can Ben change the flat bread process so bread is thick and soft?.
1...29	Ben thinks what other features would thick, soft bread have?
1...29	Ben thinks thick, soft bread would be less dense.
1...29	Ben thinks thick, soft bread might have holes or air pockets.
1...29	Ben thinks air pockets in thick, soft bread might resemble bubbles in bread.
1...30	Ben thinks Ben might create bubbles in bread by adding beer foam to dough.
1...33	Ben mixs the dough with beer foam.
1...33	Ben bakes dough.
1...36	Leo says bread is edible, thick, soft, tastes good, and not gooey.
1...37	Ben says Eureka!

Chapter 6 of [32] describes processing within this system, which illustrates learning and discovery by reasoning analogically, causal and purposive reasoning, meta-reasoning, imagination via conceptual simulation, and internal dialog between subagents in a society of mind using a language of thought. The prototype also illustrates support for semantic disambiguation, natural language constructions, metaphors, semantic domains, and conceptual blends, in communication between Tala agents. The prototype also implements these features only to a preliminary extent.

In the prototype, a Tala agent has a society of mind with subagents communicating in the Tala language, each referring to the Tala agent by a common reserved variable `?self`. Thus the TalaMind prototype simulates mental discourse (self-talk, inner speech) ([32], §2.2.4) within a Tala agent, using Tala as an interlingua.

The prototype illustrates how the TalaMind approach could support 'artificial consciousness', which is defined in section 1.5 above as a system performing observations of itself, observations of the present, the past, and potential futures, observations of its thoughts, observations of its observations, etc. Such observations are represented by expressions in the Tala natural language of thought. The prototype's illustration of this is discussed in more detail by [32] §6.3.6. [8]

The prototype demonstration system illustrated 'nested conceptual simulation' in which an agent could create nested mental contexts to represent its thoughts about other agents' actions or thoughts,

[8] A prototype routine converts Tala expressions into English text displayed by the simulation, creating some typographical errors in the output. These are just errors in a display routine, and are not errors in the internal processing of the demonstration system.

and other agents' thoughts about its actions or thoughts. This corresponds to Johnson-Laird's descriptions of iconic meta-linguistic mental models and embedding of mental models within mental models ([37], pp. 426-433).

8.3 The Discovery of Bread Simulation

Initially in this story, neither Ben nor Leo know how to make bread, nor even what bread is, nor that such a thing as bread exists. We may imagine Leo is an ancient farmer who raises goats and grows wheat grasses for the goats to eat, but does not yet know how to eat wheat himself. Ben is an ancient food and drink maker, who knows about cooking meat and making beer, perhaps from fermented wheat grass.

The discovery of bread simulation includes output from a pseudorandom 'discovery loop': After removing shells from grain Ben performs a random sequence of actions to make grain softer for eating. This eventually results either in the discovery of dough, or in making grain a "ruined mess". In the first case, Ben proceeds to discover how to make flat bread, and then leavened bread. In the second case, he says the problem is too difficult, and gives up.

Table 1 on the previous page shows a condensed example of output for the first case, omitting several less important steps in the simulation due to page limits for this paper. Each step of the form "Ben thinks …" is an internal speech act produced by a subagent of Ben communicating to another subagent of Ben, using the Tala mentalese as an interlingua. The net effect of this internal dialog is to allow Ben to perform most of the discovery of bread conceptual processing. These internal dialogs also support semantic disambiguation by Ben and Leo of each other's utterances.

Of course, it is not claimed that the story describes how humans actually discovered bread.

8.4 Limitations and evaluation of the prototype demonstration system

The ultimate research goal of the TalaMind approach is to achieve human-level artificial intelligence. Yet human-level AI involves several topics each so large that even one of them could not be addressed comprehensively within the scope of a Ph.D. thesis. The higher-level mentalities are topics for several lifetimes' research. Therefore, the TalaMind thesis [26] could not aim to prove conclusively that a system developed according to its hypotheses will achieve human-level artificial intelligence. The thesis could only present a plausibility argument for its hypotheses. Plausibility was shown by giving theoretical arguments supporting the hypotheses, and by giving arguments defending the hypotheses from potential theoretical objections, in chapters 3 and 4 of [26] and [32].

To support the plausibility argument, the thesis also discussed the prototype system illustrating how the proposed approach could work and could in principle support aspects of human-level AI, if the system and approach are fully developed – although fully developing a functional TalaMind system will need to be a long-term research effort by many researchers.

What happens at each step of the two simulated stories depends on the initial goals, knowledge, and executable concepts that Ben and Leo have within their conceptual frameworks, and depends on the conceptual processing and communication that Ben and Leo perform at each step, based on their perceptions of the environment and memories of previous events. In this respect, the story simulations are essentially predefined within the demonstration framework, even though Ben and Leo's conceptual processing involves creation of new goals, knowledge, and executable concepts.

The demonstration scenarios use pre-programmed Tala sentential concepts. In the prototype system, only the linguistic level of a TalaMind architecture is implemented, for both Tala agents. They communicate by exchanging Tala concepts, which are displayed as English sentences spoken by Ben and Leo. The story simulations are 'scripted': executable concepts work together to produce the stories, to illustrate different kinds of concept processing. ([32] §5.2)

Chapter 6, section 4 of [32] discusses how the story simulations illustrate the potential of the TalaMind approach to eventually support the higher-level mentalities of human-level AI, if the approach is fully developed in future work. This discussion is the 'evaluation of outcomes' for the story simulations. The evaluation considers how the simulations illustrate the potential of the

TalaMind approach at different steps of the stories, not just the final step of each story. Complete outputs of both story simulations are provided in [32] §6.2. A step-by-step discussion of the discovery of bread simulation is provided in Appendix B of [32].

8.5 Summary

This section has given a detailed overview of the TalaMind approach and architecture for developing a natural language of thought to support achieving human-level AI, and a detailed overview of the TalaMind prototype demonstration system.

9. What future work is needed to develop the TalaMind approach?

There is much more work needed to achieve human-level AI via the TalaMind approach, e.g.:

- Create an intelligence kernel of self-extending conceptual processes and concepts.
- Fully develop TalaMind's linguistic and archetype levels, including semantic domains and ontology.
- Integrate the linguistic level with spatiotemporal reasoning and visualization.
- Integrate an associative level, leveraging deep neural nets and Bayesian processing.
- Develop and learn ethical concepts, encyclopedic and commonsense knowledge…
- Develop higher-level mentalities including sociality, emotional intelligence, …
- Implement the above capabilities with scalability and efficiency, to support large bodies of human-level knowledge and perform human-level reasoning in real-time.
- Ensure that human-level AI is beneficial to humanity.

A longer list is given in ([32], §7.6). Clearly, there are many challenges remaining.

10. Is there an 'existence proof' for eventual success of the TalaMind approach?

Human intelligence may be an existence proof that the TalaMind approach can succeed: We know that human intelligence is based on processing in complex neural networks within the brain. We also know that humans use natural language to communicate thoughts, knowledge, and reasoning, and engage in self-talk.[9] There appears to be no evidence that human intelligence is based on formal logic. Rather, formal logic has historically been an abstract language invented by people, based on natural languages. Hence there is reason to think we should focus on developing a natural language of thought supported by processing in neural networks to achieve human-level AI, with formal logic, mathematics, and scientific notations remaining available to support human-level AI.

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 seven decades and can now envision architectures for knowledge representation using a natural language of thought to achieve human-level artificial intelligence. Such architectures may yield a 'beginning of infinity' [6] for human progress, with potential for unlimited growth of knowledge about the universe.[10]

Acknowledgements

I thank anonymous reviewers who gave constructive criticisms that led to rewriting this paper.

[9] See Fernyhough [12, 13] and §2.2.4 regarding inner speech in human intelligence.

[10] See [35] for a discussion of Deutsch's beginnings of infinity in relation to these topics in AI.

References

1. Aleksander, I., & Morton, H. (2007). Depictive architectures for synthetic phenomenology. In A. Chella & R. Manzotti (Eds.), *Artificial Consciousness*. Charlottesville, VA: Imprint Academic.
2. Ballim, A., & Wilks, Y. (1991). *Artificial believers - Ascription of belief*. Hillsdale, New Jersey: Lawrence Earlbaum Associates.
3. van Benthem, J. (2008). A brief history of natural logic. In M. K. Chakraborti, B. Löwe, M. N. Mitra, S. Sarukkai, S. (Eds.) *Logic, Navya-Nyāya & Applications*, *Homage to Bimal Krishna Matilal*. London: College Publications.
4. Carruthers, P. (1996). *Language, Thought and Consciousness – An Essay in Philosophical Psychology*. Cambridge, UK: Cambridge University Press.
5. Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 3, 200–219.
6. Deutsch, D. (2011). *The Beginning of Infinity – Explanations that Transform the World*. New York, NY: Viking Penguin.
7. Dormehl, L. (2020). Neuro-symbolic A.I. is in the future of artificial intelligence. Here's how it works. *Digital Trends*, January 5, 2020. https://www.digitaltrends.com/cool-tech/neuro-symbolic-ai-the-future/
8. Doyle, J. (1983). A society of mind – Multiple perspectives, reasoned assumptions, and virtual copies. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (pp. 309-314). San Francisco, CA: Morgan Kaufmann.
9. Drake, J. (2018). *Introduction to Logic.* Waltham Abbey, UK: ED-Tech Press.
10. Evans, V. & Green, M. (2006). *Cognitive Linguistics – An Introduction*. London, UK: Lawrence Erlbaum Associates.
11. Fauconnier, G. & Turner, M. (2002). *The Way We Think – Conceptual Blending and the Mind's Hidden Complexities*. New York, NY: Basic Books.
12. Fernyhough, C. (2016). *The Voices Within: The History and Science of How We Talk to Ourselves*. New York: Basic Books.
13. Fernyhough, C. (2017). Talking to ourselves. *Scientific American*, *317*, 2, 76–79.
14. Fodor, J.A. (2008). *LOT2 – The Language of Thought Revisited.* Oxford, UK: Oxford University Press.
15. Gärdenfors, P. (1995). Three levels of inductive inference. *Studies in Logic and the Foundations of Mathematics*, *134*, 427–449.
16. Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought.* Cambridge, MA: MIT Press.
17. Goertzel, B., & Pennachin, C., Eds. (2007). *Artificial General Intelligence*. New York, NY: Springer.
18. Goodman, N. (1983). *Fact, Fiction, and Forecast*, 4th ed. Cambridge, MA: Harvard University Press.
19. Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335-346.
20. Hart, N. (1979). *SIMON – Syntactically intelligent memory oriented networks*. Senior Thesis, Information Sciences Department, University of California at Santa Cruz. Unpublished.
21. Hitzler, P., Bianchi, F., Ebrahimi, M., & Sarker, M. K. (2019). Neuro-symbolic integration and the Semantic Web. *Semantic Web*, 0, 1-0, 1-10.
22. Hobbs, J. R. (1983). An improper treatment of quantification in ordinary English. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics* (pp. 57–63). Cambridge, MA: ACM.
23. Hudson, R. A. (2010). *An Introduction to Word Grammar*. Cambridge University Press.
24. Huyck, C. R. (2017). The neural cognitive architecture. AAAI Fall Symposium Series Technical Reports, FS-17-05, 365-370.
25. Jackendoff, R. (1989). What is a concept that a mind may grasp it? *Mind & Language*, *4*, 1 and 2, 68–102.

26. Jackson, P. C. (2014). *Toward human-level artificial intelligence – Representation and computation of meaning in natural language*. Ph.D. Thesis, Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands.
27. Jackson, P. C. (2017). Toward human-level models of minds. *AAAI Fall Symposium Series Technical Reports*, FSS-17-05, 371-375.
28. Jackson, P. C. (2018). Toward beneficial human-level AI… and beyond. *AAAI Spring Symposium Series Technical Reports*, (SS-18-01, pp. 48–53). Palo Alto, CA: AAAI Press.
29. Jackson, P. C. (2018). The intelligence level and TalaMind. *Sixth Annual Conference on Advances in Cognitive Systems*, Poster Collection, pp. 129-148.
30. Jackson, P. C. (2018). Natural language in the Common Model of Cognition. *Procedia Computer Science*, *145*, 699-709.
31. Jackson, P. C. (2018). Thoughts on bands of action. *Procedia Computer Science*, *145*, 710-716.
32. Jackson, P. C. (2019). *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language*. Mineola, NY: Dover Publications.
33. Jackson, P. C. (2019). *Introduction to Artificial Intelligence*. Third Edition. Mineola, NY: Dover Publications.
34. Jackson, P. C. (2019). I do believe in word senses. *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*. Poster Session I, pp. 321-340.
35. Jackson, P. C. (2020). Understanding understanding and ambiguity in natural language. *Procedia Computer Science*, *169*, 209–225.
36. Jackson, P. C. (2020). Toward metascience via human-level AI with metacognition. *Procedia Computer Science*, *169*, 527-534.
37. Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.
38. Kaplan, J. (2016). *Artificial Intelligence: What Everyone Needs to Know*. Oxford University Press.
39. Kartunnen, L. (2013). You will be lucky to break even. In T. H. King & V. de Paiva (Eds.) *From Quirky Case to Representing Space. Papers in Honor of Annie Zaenen*, Stanford: CSLI Publications, pp. 167–180.
40. Kartunnen, L. (2015). From Natural Logic to natural reasoning. In A. Gelbukh (Ed.), *CICLing 2015*, Part I, LNCS 9041, 295–309.
41. Katz, B. (1997). Annotating the World Wide Web using natural language. *Proceedings of the 1997 Computer-Assisted Information Searching on the Internet Conference*, 136–155. Montreal, Canada: Centre de Hautes Etudes Internationales d'Information Documentaire.
42. Kralik, J. D., Lee, J. H., Rosenbloom, P. S., Jackson, P. C., Epstein, S. L., Romero, O. J., Sanz, R., Larue, O., Schmidtke, H. R., Lee, S. W., & McGreggor, K. Metacognition for a Common Model of Cognition. (2018). *Procedia Computer Science*, *145*, 730–739.
43. Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience*, *7*, 153-160.
44. McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. In R. Chrisley & S. Begeer (Eds.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol. 2. London: Routledge Publishing. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf
45. McCarthy, J. (1959). Programs with common sense. *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 75-91. Her Majesty's Stationary Office.
46. McCarthy, J. (1992). Elephant 2000: a programming language based on speech acts. Unpublished paper. Abstract published for Keynote Address to the *22nd ACM SIGPLAN Conference on Object-Oriented Programming Systems and Applications*, 723–724.
47. McCarthy, J. (2008). The well-designed child. *Artificial Intelligence*, *172*, 18, 2003–2014.
48. McShane, M., & Nirenburg, S. A. (2012). A knowledge representation language for natural language processing, simulation and reasoning. *International Journal of Semantic Computing*, *6*, 1, 3–23.

49. Minsky, M. L. (1986). *The Society of Mind*. New York, NY: Simon & Schuster.
50. Moravcsik, J. M. (2016). *Meaning, Creativity, and the Partial Inscrutability of the Human Mind.* 2nd ed. Stanford, CA: CSLI Publications.
51. Nilsson, N. J. (2005). Human-level artificial intelligence? Be serious! *AI Magazine*, *26*, 4, Winter, 68–75.
52. Nirenburg, S., & Wilks, Y. A. (2001). What's in a symbol: Ontology, representation and language. *Journal of Experimental and Theoretical Artificial Intelligence*, *13*, 1, 9–23.
53. Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
54. Schneider, S. (2011). *The Language of Thought – A New Philosophical Direction*. Cambridge, MA: MIT Press.
55. Sowa, J. F. (1999). *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Pacific Grove, CA: Brooks/Cole Publishing.
56. Strawson, P. F. (1963). *Introduction to Logical Theory*. London: Methuen.
57. Strawson, P. F. (1950). On referring. *Mind*, *59*, 235, 320–344.
58. Thrush, T., & Winston, P. (2018). The partial mental state inducer: Learning intuition with few training examples and K-line theory. *Advances in Cognitive Systems, 7*, 97–116.
59. Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.
60. Wilks, Y. (1982). Some thoughts on procedural semantics. In W. Lehnert & M. Ringle, (Eds.), *Strategies for Natural Language Processing*. Hillsdale, NJ: Lawrence Erlbaum Associates.
61. Wilson, D., & Sperber, D. (2012). *Meaning and Relevance*. Cambridge, UK: Cambridge University Press.
62. Yudkowsky, E. (2007). Levels of organization in general intelligence. In B. Goertzel and C. Pennachin (Eds.), *Artificial General Intelligence*. Berlin: Springer-Verlag.

Paper Version Date: August 11, 2020