Proceedings of the Ninth Goal Reasoning Workshop Submitted 10/2021; published 11/2021

Toward Human-Level Goal Reasoning With A Natural Language of Thought

Philip C. Jackson, Jr. DR.PHIL.JACKSON@TALAMIND.COM
TalaMind LLC, PMB #363, 55 E. Long Lake Rd., Troy, MI, 48085 USA

Abstract

What is the nature of goal reasoning needed for human-level artificial intelligence? This research position paper contends that to achieve human-level AI, a system architecture for human-level goal reasoning 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.

1. Introduction

'Goal reasoning' is the ability of an intelligent agent to "reason about, formulate, select, and manage its goals/objectives" (Aha *et al.*, 2013). This includes all the reasoning that occurs throughout a goal's lifecycle (Roberts *et al.*, 2014) from formulation to being finished or dropped.

We are still far from creating AI systems that can perform goal reasoning as flexibly and extensively as human intelligence. The goal of this paper is to present reasons why a 'natural language of thought', supported by neural networks, would be ideal for human-level goal reasoning in systems that may eventually achieve human-level artificial intelligence. These pages will consider the following questions:

- What is human-level intelligence? What would be human-level artificial intelligence?
- What is a human-level goal? What is human-level goal reasoning?
- What is the role of natural language in human-level goal reasoning?
- What are the major options for AI understanding natural language?
- What are the major options to achieve human-level AI?
- How could a 'TalaMind' architecture (Jackson, 2014 *et seq.*) support human-level goal reasoning, with a natural language of thought?

It appears the approach to human-level goal reasoning that will be advocated in this paper has not been previously discussed in the workshops on goal reasoning that have been held since 2013. Time does not permit this paper to discuss the previous papers in the workshops in detail; the reader may wish to read previous papers about goal reasoning, such as those included in the References.

2. What is Human-Level Intelligence? What would be Human-Level Artificial Intelligence?

The issue of how to define human-level intelligence has been a challenge for AI researchers. 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 (Kaplan, 2016).

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 (Jackson, 2018).

An approach different from the Turing Test was proposed in (Jackson, 2014) and discussed in subsequent writings: 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 robustly support these capabilities, which I call *higher-level mentalities*:

- Natural Language Understanding
- 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
- Creativity and Originality
- Generality, Effectiveness, Efficiency

The higher-level mentalities together comprise a qualitative difference which would distinguish human-level AI from current AI systems and computer systems in general: While they have been topics of research,[1] no single AI system yet developed combines all of them robustly at the levels achieved by human intelligence. They are relevant specifically to this paper: The next section will discuss how human-level goal reasoning is important for the higher-level mentalities and following sections will discuss how a natural language of thought could support human-level goal reasoning in relation to the higher-level mentalities. More general discussions of the higher-level mentalities are given in (Jackson, 2019), beginning in Chapter 2 section 1.2.

Inspecting the internal design and operation of a system avoids the limitations of purely behavioral tests, like the Turing Test. People would be able to read the code for a system, inspect the internal data structures developed during a system's operation, and develop an understanding of whether a system can in principle achieve the higher-level mentalities of human-level intelligence. This is further discussed in Chapter 2, section 1.1 of (Jackson, 2019).

[1] For example, MIDCA embodies a theory of metacognition (Cox *et al.*, 2016).

3. What is a Human-Level Goal? What is Human-Level Goal Reasoning?

To support discussion, we may define a human-level goal as the grammatical object of a natural language sentence of the form "X wants Y", where the subject X refers to an agent, perhaps with human-level intelligence, and Y is any natural language grammatical object. This definition is of course very broad. It allows Y to be a nested sentence or complex phrase, perhaps involving other goals (e.g., "A wants B to want C"). The verb "wants" may be generalized to include "needs", "desires", "prefers", etc. Goals may or may not be realistic. Other grammatical forms could also be used, e.g., "X's goal is to Y".

We may define human-level goal reasoning (HLGR) as all the forms of goal reasoning that are used by the higher-level mentalities of human-level intelligence. This is a simple, open-ended definition, but as a starting point it is sufficient to support this paper's discussion. Following are some brief discussions of how goal reasoning would support several of the higher-level mentalities in a human-level AI:

- *Creativity and Originality* - Having the ability to postulate new goals, and reason about whether to pursue them, would support creativity and originality in human-level AI.
- *Self-Development and Higher-Level Learning* - A variation of the requirement for originality is a requirement for 'self-development': People can 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 reasoning about thoughts and experience to develop new methods for thinking and acting, and it includes learning by creating explanations and testing predictions using 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. To support higher-level learning, a human-level AI must have the ability to reason about, formulate, select, and manage its goals, and to reason about the goals of humans and other intelligent systems.
- *Natural Language Understanding* – Goal reasoning supports natural language understanding, which often involves reasoning about the goals that motivated natural language expressions. One may also have goals to understand what others expressed in natural language. One may reason about such goals, and decide to prioritize or abandon them, for different natural language understanding situations. Natural language supports creativity and originality in expressing new goals. As will be discussed below, virtually all forms of goal reasoning may be expressed using natural language.
- *Metacognition and Multi-Level Reasoning*[2] - 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 many different forms of meta-cognition, e.g., reasoning about reasoning, reasoning about learning, learning how to reason,

[2] For concision, (Jackson, 2014) used the term *multi-level reasoning* to refer collectively to the reasoning capabilities of human-level intelligence, including meta-reasoning, analogical reasoning, causal and purposive reasoning, abduction, induction, and deduction.

learning how to learn...[3] Goal reasoning would provide a mechanism for controlling metacognition and considering questions like “Why should I think about this?” Reasoning about goals may itself be considered a form of metacognition.

- *Imagination* – Imagination allows us to conceive things we do not know how to accomplish, and to conceive what will happen in hypothetical situations. To imagine effectively, we must know what we do not know, and then consider ways to learn what we do not know or to accomplish what we do not know how to do. A human-level AI must demonstrate imagination. Goal reasoning would provide a control mechanism for imagination.
- *Self-Awareness – Artificial Consciousness* - A human-level AI must have some degree of awareness and understanding of its own existence, its situation or relation to the world, its perceptions, thoughts, and actions, both past and present, as well as potentials for the future. So, at least some aspects of consciousness are necessary for a system to demonstrate human-level intelligence. Jackson (2014, §3.7.6) discussed how a human-level AI could perform observations that would satisfy the “axioms of being conscious” proposed by Aleksander and Morton (2007). A human-level AI would benefit from using goal reasoning with the observations that support its artificial consciousness, to be more than a passive observer of its thoughts and environment.
- *Curiosity, Self-Programming, Theory of Mind* - To support higher-level learning, an intelligent system must have another general trait, *curiosity*, which at the level of human intelligence may be described as the ability to ask relevant questions and understand relevant answers. In English, questions involve the interrogatives *who*, *what*, *where*, *when*, *why*, and *how*. The last two especially merit further discussion: A *how* question asks for a description of a method, which can be a procedure or a process. To understand the answer, an intelligent system needs to be able to represent procedures and processes, think about such representations, and ideally perform the procedures or processes described by representations, if it has the necessary physical abilities and resources. It is natural for an intelligent system to represent procedures and processes at the linguistic level of its AI architecture. With such representations it is a relatively direct step to support self-programming within an AI system. A *why* question asks for a description of either a cause or an intent. Understanding the answer requires that an intelligent system be able to support causal reasoning about physical events, and also be able to support reasoning about people’s intentions for performing actions. Reasoning about intentions involves supporting ‘Theory of Mind’, the ability for an AI system to consider itself and other intelligent agents (including people) as having minds with beliefs, desires, different possible choices, etc. Yet reasoning about intentions is essentially reasoning about goals, i.e., it is an aspect of goal reasoning.
- *Sociality, Emotions, Values* – Because humans need to discuss and share goals, goal reasoning and goal sharing are very important in human sociality. Hence the ability to reason about goals and share goals will be important for a human-level AI to support human goals and to interact socially with humans. A human-level AI will also need some understanding of

[3] Others have focused on different aspects of metacognition, such as “knowing about knowing” or “knowing about memory”. Consciousness may also be considered an aspect of metacognition.

human emotions to successfully share human goals, since emotions can drive creation, prioritization, and selection of goals. And it would need to understand and share human values if it is to achieve human-level goal reasoning and goal sharing. Research on these topics could leverage research directions that have been investigated by scientists such as Ortony et al. (1988), Ridley (1996), Picard (1997), Pinker (2002), Norman (2004), Minsky (2006), Bach (2015), Larue et al. (2018), and McDuff and Czerwinski (2018).

Human-level goal reasoning may itself be considered a higher-level mentality of human-level intelligence, since it guides and supports other higher-level mentalities as discussed above.

4. What is the Role of Natural Language in Human-Level Goal Reasoning?

Virtually all forms of human goal reasoning are expressed in human natural languages, at least when humans communicate goal reasoning to each other. There may be some forms of goal reasoning communicated with gestures or graphical signs, yet arguably any such goal reasoning could also be communicated in natural language.

This observation suggests we do not need to know how the human brain represents goal reasoning internally, at least for the purpose of this paper's discussion of how human-level goal reasoning could be supported by a natural language of thought. It may suffice to consider the forms of goal reasoning that humans can express using natural language.

Thus, all the forms of goal reasoning previously discussed in AI research papers (at goal reasoning workshops since 2013) have been expressed in natural language, as well as in several different symbolic representations used in the research papers.

And tautologically, natural language could be used to express any forms of goal reasoning not previously discussed by AI researchers that they could express and discuss in natural language in future papers.

Thus, understanding and representing human-level goal reasoning may be considered as a subproblem of understanding and representing the semantics of human natural languages, in developing human-level artificial intelligence.

NOTE: The following sections 5 through 7.1 discuss general questions related to natural language and human-level AI to provide a foundation for section 7.2, which discusses representation of goals and human-level goal reasoning with a natural language of thought. Readers who are only interested in goal reasoning may wish to jump to section 7.2 and read the intervening sections later, if needed.

5. What are the Major Options for AI Understanding Natural Language?

Understanding natural language was listed in section 2 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.

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

One 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.

In AI research, 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.

Although this has been the traditional approach, it is not the approach advocated by this paper, for reasons discussed in section 6.1 below.

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

This paper advocates a major alternative for symbolic processing of natural language, in combination with neural networks. This 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. (Jackson, 2019)

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 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. (Jackson, 2019) And it involves more than representations to annotate meaning of natural language expressions (e.g., Van Gysel *et al*., 2021; Banarescu *et al*., 2013). The TalaMind approach envisions annotating and using natural language expressions within an AI system, as representations of thoughts.

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 (Jackson, 2019, pp. 156-177).

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. More discussion of this is given in section 7 of (Jackson, 2021).

6. What are the Major Options to Achieve Human-Level Artificial Intelligence?

Logically, there are three major alternatives toward this goal, discussed in the following subsections:

- Purely symbolic approaches to HLAI.
- Neural network architectures.
- Hybrid systems

6.1 Purely Symbolic Approaches to HLAI

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

Based on the generality of symbolic logic, one might think it should be possible to achieve human-level AI by developing systems which only use symbolic logic, with extensions of first-order logic, along with other symbolic approaches, such as semantic networks and frame-based systems.

Such approaches have been developed in AI research for many years. However, symbolic logic effectively handicaps achieving human-level AI, because it is not as flexible as natural language for representing human thoughts and knowledge: Formal logic specializes and standardizes the use of certain natural language words and phrases, and in principle, anything that can be expressed in formal logic could be translated into equivalent expressions in natural language.

But the opposite is not true: Natural language can express ideas and concepts much more flexibly than formal logic. Natural language allows communication without needing to be precise about everything at once (Sowa, 2007). Natural language supports expressing thoughts about what you think or think other people think, thoughts about irrational or self-contradictory situations, emotions, etc.

To achieve human-level artificial intelligence, a system will need the ability to understand the full range of human thoughts that can be expressed in a natural language like English. No existing formal logic language can represent this range of thoughts. A natural language like English can already do this, perhaps as well as any artificial, formal logic language ever could. The development of system architectures for human-level artificial intelligence (including human-level goal reasoning) would be greatly handicapped by relying only on formal logic systems.

So, the TalaMind approach (Jackson, 2014 *et seq.*) advocates representing and using a natural language like English as a 'natural language of thought' within AI systems that may eventually achieve human-level artificial intelligence.

6.2 Neural Network Architectures for HLAI

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 (Huyck, 2017). 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 (Markram, 2006).

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 childlike learning process, and avoid complete dependence on neural nets by allowing an AI system to use other computational methods when neural nets are not needed.

6.3 Hybrid Architectures for HLAI

It should be possible to develop hybrid (‘neuro-symbolic’) architectures, combining symbolic processing and neural networks to support eventually achieving human-level AI. Such architectures could have substantial advantages: Symbolic processing would support representing, reasoning, and learning with sentential structures, networks, contexts, etc. Neural networks would support learning, representing, and recognizing complex patterns and behaviors that are not easily defined by symbolic expressions.

This paper will advocate a class of hybrid architectures called the ‘TalaMind architecture’ (Jackson, 2014) and will focus on discussing the symbolic processing side of the architecture, to support a natural language of thought. Integration of neural networks is a topic for ongoing and future research.

7. How Could a 'TalaMind' Architecture Support Human-Level Goal Reasoning, with a Natural Language of Thought?

7.1 An Overview of the TalaMind Approach to Human-Level Artificial Intelligence

The TalaMind approach was proposed by (Jackson, 2014) for research toward eventually achieving human-level artificial intelligence. 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 (Evans & Green, 2006) (Fauconnier, 1994).

The first hypothesis essentially describes the 'seed AI' approach in AGI (Yudkowsky, 2007). 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.

To support developing systems according to these hypotheses, (Jackson, 2014) proposed the TalaMind architecture having three levels of conceptual representation and processing, called the linguistic, archetype, and associative levels. (See Figure 1, next page.)

These levels were adapted from Gärdenfors' (1995) paper on levels of inductive inference. He called them 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.

At the linguistic level, the architecture includes a natural language of thought called Tala. The linguistic level also includes 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, expressed in Tala.

In the TalaMind[4] approach, conceptual processes can be implemented with 'executable concepts', also expressed in Tala, which can create and modify executable concepts (Jackson, 2019, pp. 214-217). The potential scope of conceptual processes would be computationally universal (*ibid*, p.73).

[4] TalaMind® and Tala® are trademarks of TalaMind LLC, to support future development.

At the archetype level, cognitive categories and concepts may be represented using methods such as conceptual spaces, image schemas, semantic frames, radial categories, etc.

The associative level would typically interface with a real-world environment and support deep neural networks, Bayesian processing, etc. At present, the TalaMind approach does not prescribe specific research choices at the archetype and associative levels.

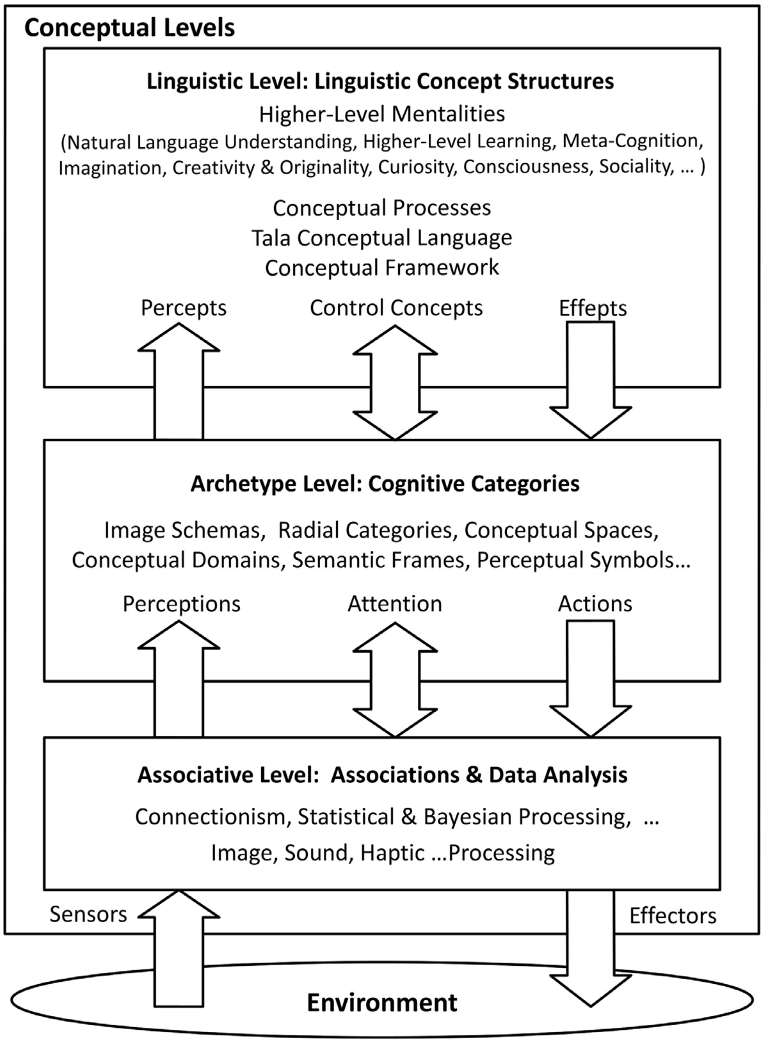
Conceptual Levels
Linguistic Level: Linguistic Concept Structures
Higher-Level Mentalities
(Natural Language Understanding, Higher-Level Learning, Meta-Cognition, Imagination, Creativity & Originality, Curiosity, Consciousness, Sociality, …)
Conceptual Processes
Tala Conceptual Language
Conceptual Framework
Percepts
Control Concepts
Effepts
Archetype Level: Cognitive Categories
Image Schemas, Radial Categories, Conceptual Spaces, Conceptual Domains, Semantic Frames, Perceptual Symbols…
Perceptions
Attention
Actions
Associative Level: Associations & Data Analysis
Connectionism, Statistical & Bayesian Processing, …
Image, Sound, Haptic …Processing
Sensors
Effectors
Environment

Figure 1. TalaMind System Architecture

Thus, a fully developed TalaMind system would not be a purely symbolic system nor a purely neural system. It would be neuro-symbolic. At its linguistic level, it would have symbolic data structures representing natural language expressions, word senses, contexts, etc. Scientists would be able to inspect these data structures and the operation of the TalaMind system using computer software tools, in the same way that computer programmers can inspect data structures being developed and used by computer programs in general. Inspection would not be limited to interacting with the AI system using natural language, in a Turing Test.

The TalaMind 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. So, the TalaMind architecture is actually a broad class of architectures, open to further design choices at each level. For concision, a system with a TalaMind architecture is called a 'Tala agent'.

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 (2002) 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. 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.

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. The topic of 'primitive' words for the Tala language is discussed in (Jackson, 2019, pp. 125-128).

Chapter 5 of (Jackson, 2019) presents a design for the syntax of the Tala conceptual language, which is general and flexible and covers many of the issues discussed by Hudson (2010) 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 will be a large effort that could occupy multiple researchers. Chapters 5 and 6 (*ibid*) discuss the TalaMind prototype demonstration system.

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 may not involve disambiguation of word senses because Tala expressions can include pointers to word senses and referents.

The Tala syntax also supports nongrammatical natural language expressions because people frequently use nongrammatical language, and a human-level AI needs to be able to represent and try to understand whatever people say. A Tala agent should be able to represent and understand metaphors, metonyms, anaphora, idioms, multiple negatives, etc. The Tala syntax is somewhat nonprescriptive, open, and flexible, for example by making parts of speech optional. It should be clear that Tala is not a 'controlled natural language'.

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 (1986). This broader, generalized sense corresponds to a paper by Doyle (1983), who referred to a multiagent system using a language of thought for internal communication, though Doyle did not discuss a 'natural language of thought'.

7.2 The TalaMind Approach to Representation of Goals and Human-Level Goal Reasoning

In considering this topic, it's appropriate to consider both the discussion of goal representation and processing in the 'TalaMind thesis' (Jackson, 2014), and the more general potential of the TalaMind approach to support human-level goal reasoning.

7.2.1 Representation of Goals and Limited Goal Reasoning in the TalaMind Demonstration

The focus of the TalaMind thesis is on presenting a research direction toward eventual creation of human-level artificial intelligence. The thesis discusses representation and processing of goals only to a limited extent, sufficient to support the more general discussion. The thesis does not discuss goal reasoning *per se*, though parts of the prototype demonstration illustrate the potential of the TalaMind approach to support goal reasoning.

In the demonstration, a goal is a Tala conceptual structure using the verb "want". The object of a goal is itself a Tala expression. For example, a goal might be:

```
(want (wusage verb)
   (subj ?self)
   (obj
      (examine (wusage verb)
          (subj ?self)
          (obj (grain (wusage noun)]
```

This goal says in effect "I want to examine grain", though it does not use an infinitive. (When the goal is displayed in output by the system, the Tala agent's name is substituted for `?self`.)

Within the conceptual framework of a Tala agent, the `(goals)` slot stores the agent's current goals, within its perceived reality. Goals can also be represented and apply only within contexts being considered by a Tala agent.

The TalaMind demonstration system is a functional prototype in which two Tala agents, named Ben and Leo, interact in a simulated environment by exchanging Tala concepts and by performing actions represented by Tala concepts. Each Tala agent has its own TalaMind conceptual framework and conceptual processes.

To the human observer, a simulation is displayed as a sequence of English sentences, in effect a story, describing interactions between Ben and Leo, their actions and percepts in the environment, and their thoughts. The story that is simulated depends on the initial concepts that Ben and Leo have, their initial percepts of the simulated environment, and how their executable concepts process their perceptions to generate goals and actions, leading to further perceptions and actions at subsequent steps of the story. All of the concepts in the simulation are predefined.

In the 'discovery of bread' TalaMind demonstration, the following pre-defined Tala sentences are used for representation and limited reasoning about goals:

```
Leo wants Ben to make edible grain.
Leo knows that if he wants to do X but cannot do X himself, then he could
want someone else to do X.
Leo knows that if he wants someone to do X, he should ask them to do X.
Ben wants Ben to know whether humans perhaps can eat grain.
Ben wants Ben to know how Ben can make grain be food for people.
Ben wants Ben to experiment with grain.
Ben wants Ben to examine grain.
If I want to experiment with X Then ask Leo to turn over some to me for
experiments.
If A asks me to give X to A Then If I want A to make X edible And I have
excess X Then give X to A
If someone gives me X And I want to examine X Then examine X.
If X resembles Y And I want to know if X is edible And Y is edible Then
imagine an analogy from Y to X focused on food for people.[5]
If remove S from X must precede eating X And I want to know how to make X be
food for people And X resembles Y Then imagine an analogy from Y to X
focused on removing S.
If I think perhaps people would prefer eating thick soft X over eating flat
X Then think how can I make thick soft X? And want to make thick soft X.
Ben wants Ben to make thick, soft bread.
If I think A asks can I give more X to A for experiments Then If I want A to
make X edible And X is in current-domains And I have excess X Then Give more
X to A.
```

Within the prototype system these sentences are represented as syntactic list structures, like the *“I want to examine grain”* example above. The system’s ‘FlatEnglish’ routine displays the sentences as shown above, to make the demonstration more easily understandable for people.

The TalaMind prototype logic supports creation of goals and pattern-matching of goals with executable concepts (‘xconcepts’),[6] and automatically detects when goals have been satisfied. The logic also automatically deletes goals that have been satisfied and prevents attempting to process goals that are already being processed by other xconcepts. However, the logic does not automatically propagate satisfaction between goals, nor automatically retract goals if they are no longer needed. This functionality was not implemented in the prototype since it has been studied in previous research. It could be an enhancement in a future version of the system.

[5] The logic for this is actually more general, being written to match any action that an agent can perform on Y that a Tala agent wants to perform on X.

[6] An executable concept is a concept that describes a process or behavior that may be performed by a Tala agent, i.e., a sequence of steps to perform, conditions and iterations, etc. The steps to perform may include assertions or deletions of concepts in the conceptual framework, including creation and modification of other executable concepts. Conditions may include tests on percepts, goals, finding concepts within the conceptual framework, etc. Pattern-matching may be used to express conditions, so that an executable concept may process all or part of a Tala concept.

7.2.2 Potential of the TalaMind Approach to Support Human-Level Goal Reasoning

Considering the discussions above and in previous sections, we can now discuss the potential for a natural language of thought to support human-level goal reasoning in systems developed following the TalaMind approach.

Section 3 gave an initial definition and discussion of human-level goal reasoning as the forms of goal reasoning that are used by the higher-level mentalities of human-level intelligence. Section 3 noted that natural language plays a central role in enabling the higher-level mentalities.

Section 4 noted that virtually all forms of goal reasoning are expressed in human natural languages, when humans communicate goal reasoning to each other.

The TalaMind demonstration system (discussed in the previous section) illustrated how a natural language of thought could support representation and reasoning about goals, although the demonstration was very limited.

When fully developed, a TalaMind system would support the unconstrained syntax of English (and potentially other natural languages) in a natural language of thought. The natural language of thought could enable the TalaMind system to support human-level representation of thoughts about goals. For example, a human-level AI could represent and reason about the following thoughts involving goals, using a natural language of thought:

```
Goal A conflicts with goal B, and supports goal C.
My goal for today is to plan my goals for this month.
I want X, hope for Y, need at least Z, and may only achieve W.
The two parties have compatible / conflicting goals.
X wants to have his cake and eat it, too.
Why does X want to learn Y’s goals?
Is it realistic to imagine ways to achieve goal C?
What goals should I have before and after achieving goal X?
Can redefining goal C make it more worthwhile or easier to achieve?
Should I set goals that are easy to achieve, or goals that are difficult?
```

In theory, any goal reasoning thoughts that humans can express in natural language would be possible for a TalaMind system to represent in a natural language of thought. This would provide a foundation for the system to achieve human-level goal reasoning, and support human-level artificial intelligence.

8. A Limitation of AI Systems for Natural Language Semantics of Goals

However, there is at least one limitation that should be kept in mind: Representing syntax is not the same as understanding semantics, especially when semantics involves human subjective concepts such as emotions. So, for example, if an AI system represents and processes the sentence *“My goal is for John to be happy”*, its ability to understand the sentence will be limited by the extent to which it cannot understand (and perhaps, subjectively feel) human happiness.

This limitation affects all AI systems, not just systems following the TalaMind approach. It is not specific to the use of a natural language of thought. It underscores the point stated in section

3, that a human-level AI will need some understanding of human emotions and human values to successfully share human goals, and to achieve human-level goal reasoning.

More generally, a human-level AI will need an understanding of ethical values, rules, and principles to share human goals, and to support 'beneficial AI' – AI that is beneficial to humanity and to life in general (Bringsjord, Arkoudas, & Bello, 2006; Tegmark, 2017). Future issues related to beneficial AI are discussed in more detail by (Jackson, 2019, §8.2).

9. Summary

A 'natural language of thought' would be ideal for support of human-level goal reasoning in systems that may eventually achieve human-level artificial intelligence. The TalaMind approach envisions a neuro-symbolic architecture for such systems, which will integrate processing at linguistic, archetype, and associative levels.

Of course, there is much more work needed to achieve human-level AI, including human-level goal reasoning, via the TalaMind approach. Ideally, future research will focus on developing Tala agents that perform goal-reasoning to support the higher-level mentalities of human-level intelligence, including creativity, metacognition, imagination, artificial consciousness, curiosity, and sociality, as well as natural language understanding. In principle, future research should include empirical studies in domains where performance of Tala agents could be compared with baselines achieved by humans or other AI systems.

Acknowledgements

I thank Sravya Kondrakunta for inviting submission of a paper with this scope to the Ninth Goal Reasoning Workshop. Two anonymous reviewers gave constructive comments and questions that led to changes which hopefully improved the paper's clarity and completeness.

References

Aha, D. W., Cox, M. T., Muñoz-Avila, H. (2013) Preface. *Goal Reasoning: Papers from the ACS Workshop*, Baltimore, MD. https://drum.lib.umd.edu/handle/1903/14740

Aleksander, I., Morton, H. (2007) Depictive architectures for synthetic phenomenology. In Chella & Manzotti (2007), pp. 67-81.

Bach, J. (2015) Modeling motivation in MicroPsi 2. *Artificial General Intelligence, 8th International Conference*, AGI 2015, pp. 3-13.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N. (2013) Abstract Meaning Representation for Sembanking. *LAW@ACL*.

Bengfort, B., Cox, M. T. (2015) Interactive knowledge-goal reasoning. *Goal Reasoning: Papers from the ACS Workshop*. http://www.cc.gatech.edu/~svattam/goal-reasoning

Bobrow, R., Brinn, M., Burstein, M., Laddaga, R. (2013) Goal substitution in response to surprises. *Goal Reasoning: Papers from the ACS Workshop*, Baltimore, MD.

Bringsjord, S., Arkoudas, K., Bello, P. (2006) Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, July 2006, 38-44.

Chella, A., Manzotti, R., eds. (2007) *Artificial Consciousness*. Imprint Academic.

Cox, M. T. (2013) Question-based problem recognition and goal-driven autonomy. *Goal Reasoning: Papers from the ACS Workshop*, Baltimore, MD.

Cox, M. T., Alavi, Z., Dannenhauer, D., Eyorokon, V., Munoz-Avila, H., Perlis, D. (2016) MIDCA: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy. Proceedings of the 30th AAAI Conference on AI (AAAI-16), 3712-3718.

Doyle, J. (1983) A society of mind – multiple perspectives, reasoned assumptions, and virtual copies. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 309-314.

Epstein, S. L., Korpan, R. (2020) Metareasoning and path planning for autonomous indoor navigation. *Integrated Execution (IntEx) / Goal Reasoning (GR) Workshop, ICAPS 2020*.

Evans, V., Green, M. (2006) *Cognitive Linguistics – An Introduction*. Lawrence Erlbaum Associates.

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

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

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

Goertzel, B., Pennachin, C. (2007) *Artificial General Intelligence*. Springer Publishing

Hudson, R. (2010) *An introduction to Word Grammar*. Cambridge University Press.

Huyck, C. R. (2017) The neural cognitive architecture. *AAAI Fall Symposium Series Technical Reports*, FS-17-05: 363-370

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

Jackson, P. C. (2018). Toward beneficial human-level AI… and beyond. *AAAI Spring Symposium Series Technical Reports*, SS-18-01, 48-53.

Jackson, P. C. (2019) *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language*. Dover Publications.

Jackson, P. C. (2020) Understanding understanding and ambiguity in natural language. *Procedia Computer Science*, 169: 209-225.

Jackson, P. C. (2021) On achieving human-level knowledge representation by developing a natural language of thought. *Procedia Computer Science*, 190: 388-407.

Kaplan, J. (2016) *Artificial intelligence – what everyone needs to know*. Oxford University Press.

Larue, G., West, R., Rosenbloom, P.S., Dancy, C.L., Samsonovich, A.V., Petters, D., Juvina, I. (2018) Emotion in the Common Model of Cognition. *Procedia Computer Science*, 145, 740-746.

Markram, H. (2006) The Blue Brain project. *Nature Reviews Neuroscience*, 7:153-160.

Minsky, M. L. (1986) *The Society of Mind*. Simon & Schuster.

Minsky, M. L. (2006) *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster.

McDuff, D., Czerwinski, M. (2018) Designing emotionally sentient agents. *Communications ACM*, 61, 12, 74-83.

Mohammad, Z., Cox, M. T. (2020) Rebel agents that adapt to goal expectation failures. Multi-agent goal recognition as implicit ad-hoc teamwork. *Integrated Execution (IntEx) / Goal Reasoning (GR) Workshop, ICAPS 2020.*

Norman, D. A. (2004) *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books.

Ortony, A., Clore, G. L., Collins, A. (1988) *The Cognitive Structure of Emotions*. Cambridge University Press.

Picard, R. (1997) *Affective Computing*. MIT Press.

Pinker, S. (2002) *The Blank Slate: The Modern Denial of Human Nature*. Penguin Books.

Ridley, M. (1996) *The Origins of Virtue*. Viking.

Roberts, M., Vattam, S., Alford, R., Auslander, B., Karneeb, J., Molineaux, M., Apker, T., Wilson, M., McMahon, J., Aha, D. (2014) Iterative goal refinement for robotics. In *Working Notes of the Planning and Robotics Workshop at ICAPS*. AAAI.

Sowa, J. F. (2007) Fads and fallacies about logic. *IEEE Intelligent Systems*, March 2007, 22:2, 84-87.

Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*. Alfred A. Knopf.

Vattam, S., Klenk, M., Molineaux, M. Aha, D. W. (2013) Breadth of approaches to goal reasoning: A research survey. *Goal Reasoning: Papers from the ACS Workshop*, Baltimore, MD.

Van Gysel, J. E. L., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C., Hajic, J., Martin, J. H., Oepen, S., Palmer, M., Pustejovsky, J., Vallejos, R., Xue, N. (2021) Designing a uniform meaning representation for natural language processing. *Künstliche Intelligenz.*

Wright, B. (2020) Multi-agent goal recognition as implicit ad-hoc teamwork. *Integrated Execution (IntEx) / Goal Reasoning (GR) Workshop, ICAPS 2020.*

Yudkowsky, E. (2007) Levels of organization in general intelligence. In Goertzel & Pennachin (2007), pp. 389-501.