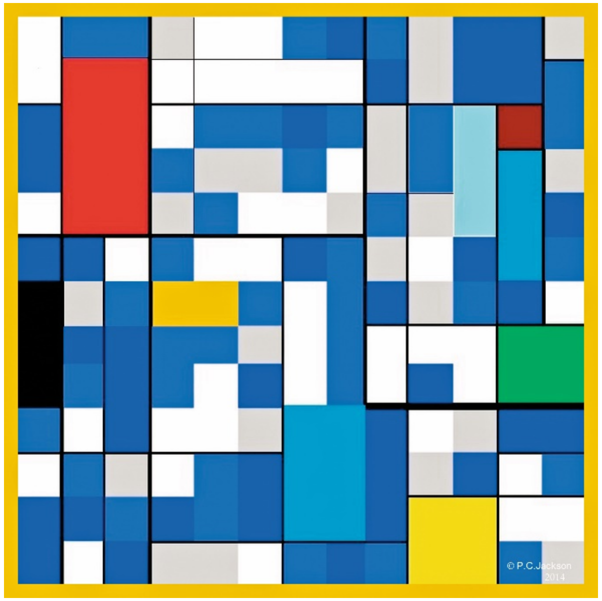
Toward Human-Level Goal Reasoning with a Natural Language of Thought

Philip C. Jackson, Jr.
TalaMind LLC



Goal Reasoning Workshop Presentation
November 15, 2021

Jackson, P. C. (2021) Toward human-level goal reasoning with a natural language of thought. *Proceedings of the Ninth Goal Reasoning Workshop*.

The Goal for This Talk about Goal Reasoning:

To persuade you that human-level AI systems should use a natural language of thought to achieve human-level goal reasoning.

To hopefully achieve this goal, I'll discuss:

- What is human-level intelligence and what would be human-level AI?
- What is a human-level goal and what would be human-level goal reasoning?
- What is the role of natural language in human-level goal reasoning? How can we represent human-level goals?
- What are the major options for AI understanding natural language and achieving human-level AI?
- What would be a natural language of thought in an AI system?
- How could a natural language of thought support human-level AI and human-level goal reasoning? Why is symbolic logic insufficient?

What Is Human-Level Intelligence?

Defining '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.

A Turing Test may help recognize human-level AI if it is created, but does not define intelligence nor indicate how to achieve human-level AI.

Also, a Turing Test focuses on recognizing human-identical AI, indistinguishable from humans.

It may be sufficient (and even important, for beneficial AI) to develop systems that are <u>human-like</u>, and <u>understandable by humans</u>, rather than human-identical.

An approach different from the Turing Test was proposed in [Jackson, 2014]:

> Define human-level intelligence by identifying abilities achieved by humans and not yet achieved by any AI system. I call these *higher-level mentalities*.

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

Higher-Level Mentalities of Human-Level Intelligence

The *higher-level mentalities* comprise a qualitative difference that would distinguish human-level AI from current AI systems and computer systems in general:

- 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

While they have been topics of research*, no single AI system yet developed combines all of them at the levels achieved by human intelligence.**

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

** Artificial consciousness does not imply an AI system would have the human subjective experience of consciousness.

What is a Human-Level Goal? What would be HLGR?

I would define human-level goals as **goals involved in achieving the higher-level mentalities of human-level intelligence**:

- 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

These mentalities support human-level goals related to employment, social relations, entertainment, etc.

Of course, humans have other, important goals related to biological survival, which help motivate higher-level human goals.

I would 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.**

Examples of Human-Level Goal Reasoning, …

- Creativity and Originality – **The ability to postulate new goals, and reason about whether to pursue them, will support creativity and originality in human-level AI.**

- Self-Development and Higher-Level Learning - includes reasoning about thoughts and experience to develop new methods for thinking and acting and **includes learning by creating explanations and testing predictions using causal and purposive reasoning**.

  The term 'higher-level learning' distinguishes these from lower-level forms of learning studied in much research on machine learning, e.g., neural networks.

  Purposive reasoning is reasoning with and about goals. So, higher-level learning involves goal reasoning.

- Natural Language Understanding – **often involves reasoning about the goals that motivated natural language expressions. One may also have goals to understand what others expressed in natural language.**

Examples of Human-Level Goal Reasoning, …

- Metacognition - **Goal reasoning could 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 – **Goal reasoning could provide a control mechanism for imagination** in human-level AI**,** and considering questions like **"Why should I imagine this?"**

  Imagination includes thinking about things we do not know how to accomplish and thinking about what will happen in hypothetical situations.

- Self-Awareness – Artificial Consciousness - **Goal reasoning could enable a human-level AI to be more than a passive observer of its thoughts and environment**.*

---

* 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), i.e., *Observation of an external environment. Observation of oneself 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.*

…Examples of Human-Level Goal Reasoning

- Curiosity – may be described as the ability to ask relevant questions and understand relevant answers. A “why” question asks for a description of either a cause or an intent. Understanding intent requires that an intelligent system be able to support reasoning about people’s intentions for performing actions. **Reasoning about intentions is essentially reasoning about goals.**

- Sociality, Emotions, Values – A human-level AI will need **some understanding of human emotions to successfully share human goals, since emotions can drive creation and prioritization of goals**.

- **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.

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

- **Essentially all forms of human goal reasoning are expressed by people in natural languages, at least when humans communicate goal reasoning to each other.** *

  This suggests we may not need to know how the human brain represents goal reasoning internally, at least for discussion of how human-level goal reasoning could be supported by a natural language of thought.

- If we take this approach, then 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.

* There may be forms of goal reasoning communicated with gestures or graphical signs, but arguably any such goal reasoning could also be communicated in natural language.

What Are The Major Options for Natural Language Understanding?

The major options for NLU parallel the major approaches toward achieving human-level AI:

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

One can argue theoretically that symbolic and neural network approaches could each be successful, based on computational generality.

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

1. **Treating natural language as external data**.
   - Translate natural language to and from some simpler internal symbolic language.
     (predicate calculus, frame-based languages, conceptual graphs, n-tuples, etc.)
   - This has been the traditional AI approach for many decades, yet it is still far from achieving human-level natural language understanding.

2. **Using natural language as a 'language of thought' within an AI system**.

What would be a ‘Natural Language of Thought’ in an AI system?

In this approach, thoughts are represented by natural language data structures.

Inference and conceptual processing are performed with these natural language data structures.

Other symbolic languages would support implementation, pattern-matching, and interpretation of natural language data structures.

This involves more than representing and using syntax of natural language to represent thoughts: It also involves representing and using semantics of natural language, to represent thoughts. (Jackson, 2014 *et seq*.)

It involves more than annotating natural language expressions to represent meaning. It involves annotating and using natural language expressions within an AI system, as representations of thoughts.

There does not appear to be any valid theoretical reason why a natural language like English cannot be used in this way by an AI system for its language of thought.

The TalaMind thesis (Jackson, 2014) advocates this approach toward achieving human-level AI.

Limitations of Symbolic Logic for Understanding NL

In principle, anything that can be expressed in formal logic could be translated into equivalent expressions in natural language.

Yet natural language can express ideas and concepts much more flexibly than formal logic.

Natural language supports expressing thoughts about what you think or think other people think, thoughts about irrational or self-contradictory situations, emotions, etc.

Natural language allows expressing thoughts without needing to be precise about everything at once (cf. Sowa, 2007).

Limitations of Neural Networks for Achieving HLAI

Switching the discussion to neural networks:

There does not appear to be any theoretical reason in principle that prevents neural networks from achieving a fully general human-level AI.

However, human neurons are much more complex than 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 emulate human neurons in such orders of magnitude with a computer system, anytime soon.

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.

This suggests that **research on neural networks to achieve human-level AI should be pursued in conjunction with symbolic methods that support explanations in a natural language like English.**

Hybrid Architectures for Achieving HLAI

It should be possible to develop such 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 and recognizing complex patterns and behaviors that are not easily represented by symbolic expressions.

I advocate a class of hybrid architectures called the ‘TalaMind architecture’ (Jackson, 2014 *et seq*.).

I have focused 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.

TalaMind Architecture Research Direction

If fully developed, after future research:

At the linguistic level, TalaMind will:

- have a language of thought (called ‘Tala’) with the unconstrained syntax and semantics of English.
- support reasoning directly with the syntax and semantics of English.

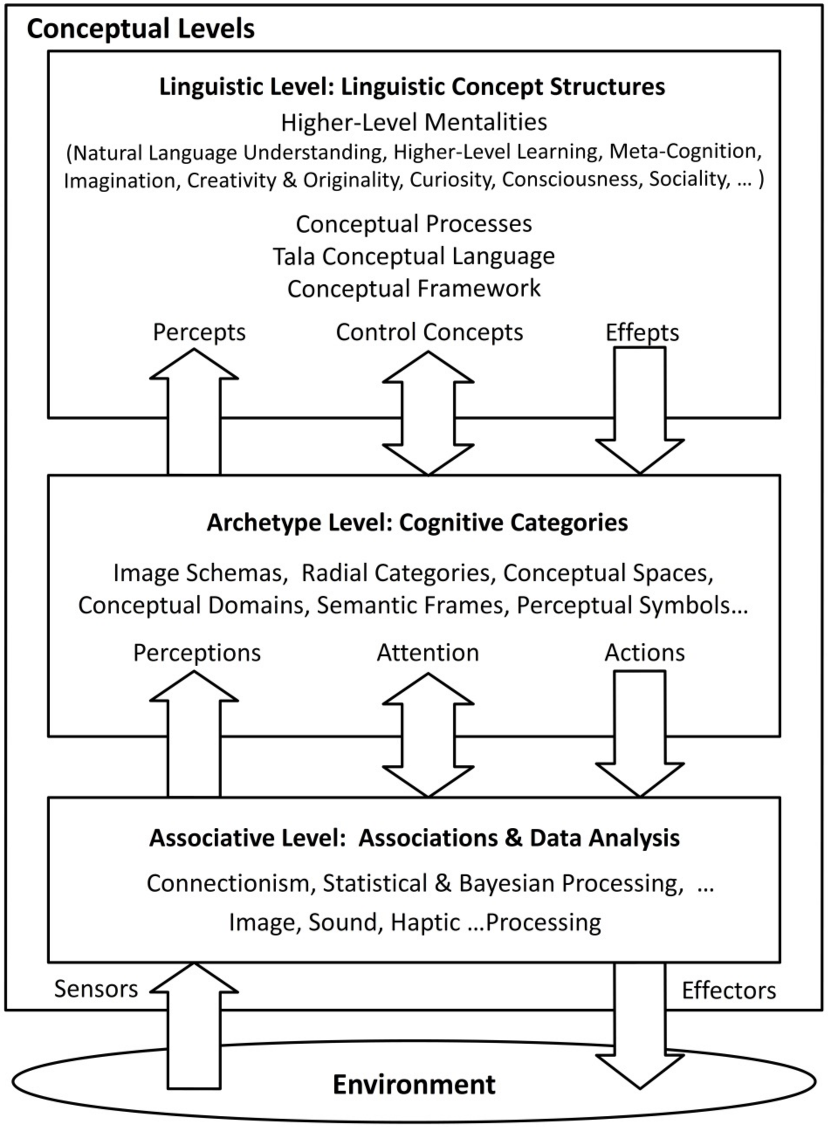
Tala (the natural language mentalese):

- will be understandable to humans and open to human inspection

These features will be important for achieving human-level AI.

The thesis prototype system illustrates these features, with limitations.

For concision, a system with a TalaMind architecture is called a ‘Tala agent’.


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

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

Representation of Goals in TalaMind Thesis

The ‘TalaMind thesis’ (Jackson, 2014) discusses representation and processing of goals only to a limited extent, sufficient to support the more general discussion. It does not discuss goal reasoning *per se*, though the prototype illustrates potential 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 “I want to examine grain”, though it does not use an infinitive.

The syntax allows goals to be a nested sentence or complex phrase, perhaps involving other goals (e.g., **“A wants B to want C”**).

Representation of Goals in TalaMind Thesis

The following sentences are used for representation and limited reasoning about goals, in the 'discovery of bread' TalaMind demonstration:

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.

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.

Potential of TalaMind Approach for HLGR

In principle, 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.

A Limitation of AI Systems for NL 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 applies to formal logic, and in general to representation of thoughts in AI systems.

So, 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.

Within an AI system, a representation of emotions could help guide choices of goals. These are topics for future research in developing the TalaMind approach. (cf. Jackson, 2019, pp. 27-28)

Summary

The 'hard truth': Symbolic logic handicaps achieving human-level AI. A 'natural language of thought' would be ideal to support human-level goal reasoning and eventually achieve human-level artificial intelligence, although no approach to HLAI will be easy.

The TalaMind approach envisions a neuro-symbolic architecture for HLAI, 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.

"Better by far to embrace the hard truth than a reassuring fable.
If we crave some cosmic purpose, then let us find ourselves a worthy goal."*

* Carl Sagan, Ann Druyan (1994) *Pale Blue Dot: A Vision of the Human Future in Space,* p.54, Ballantine Books.

Questions?

References

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

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*.

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

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.

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

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

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*.