ACS Poster Collection (2018) 1−20 Published 8/2018; Paper 8

The Intelligence Level and TalaMind

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

Abstract

This paper discusses theoretical and practical limitations for Newell's (1982) definition of the 'knowledge level'. An alternative definition is given for an 'intelligence level', corresponding to human-level artificial intelligence. Topics for research and development in the intelligence level are discussed, illustrated by the TalaMind approach to human-level AI (Jackson, 2014).

1. Introduction

In the decades after their groundbreaking research on artificial intelligence in the 1950's, Allen Newell and Herbert Simon continued their research and wrote a series of papers about cognitive systems. Simon wrote several books, including one with Newell in 1972 on *Human Problem Solving*. Newell wrote a book on *Unified Theories of Cognition*, published in 1990.

This paper focuses on three works: Newell and Simon's (1976) paper on physical symbol systems, Newell's (1982) paper on the 'knowledge level', and Newell's (1990) book.

After noting an important ability which does not occur intrinsically for physical symbol systems, important theoretical and practical limitations for Newell's (1982) definition of the knowledge level are discussed. Newell (1990) continued to advocate this problematic definition of the knowledge level, and extended it to give a counterintuitive definition of 'perfect intelligence'.

This paper presents an alternative definition for an 'intelligence level' of computer systems, corresponding to human-level artificial intelligence. Several topics for research and development in the intelligence level are discussed.

2. Review of Previous Research

2.1 Newell and Simon's 1976 Paper on 'Symbols and Search'

This paper was Newell and Simon's Turing Lecture for the 1975 Turing Award. To consider artificial intelligence and computer science in general, they discussed the importance of *'laws of qualitative structure'* in characterizing knowledge about scientific domains. As examples of such laws they cited the cell doctrine in biology, plate tectonics in geology, the germ theory of disease, and the doctrine of atomism in chemistry.

They then described 'physical symbol systems' as a qualitative structure in computer science, and hypothesized a law of qualitative structure for intelligent systems, the *Physical Symbol System Hypothesis (PSSH)*:

> "A physical symbol system has the necessary and sufficient means for general intelligent action."

Briefly, they defined physical symbols as physical patterns, which can occur in expressions (symbol structures). A system can contain a collection of expressions, and processes which operate on the expressions to produce new expressions. Expressions can designate processes to perform. The system can interpret expressions to perform the processes they designate.

Newell and Simon noted their definition of a physical symbol system essentially describes the symbolic processing abilities of digital computers. They wrote "By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action." So PSSH essentially hypothesizes that digital computers can support human-level artificial intelligence.

Newell and Simon said that a physical symbol system can have expressions which designate "external objects", but they did not discuss designation of external entities that exist mainly via shared understanding among humans, e.g. money, corporations, laws, etc. (Harari, 2015) While they hypothesized symbolic systems could achieve human-level AI, they did not mention what everyone knew was the *status quo*, that computers had to be programmed by humans to successfully designate such external entities and support or perform the processes people intended them to perform.

Computer programs did not "understand" what their symbols designated and what they were intended to do in the human world well enough for programs to debug themselves and perform as people intended. Such an understanding is far from being intrinsic to systems at the symbolic level. I will say more, later in this paper, about what such an understanding would entail.

2.2 Newell's 1982 Paper *'The Knowledge Level'*

In this paper, Newell asked the following questions: "What is the nature of knowledge? How is it related to representation? What is it that a system has, when it has knowledge?"

2.2.1 The Nature of Computer Systems Levels

Before answering his questions, Newell reviewed the nature of computer system levels previously developed, e.g. the electronic device level, circuit level, logic level, register-transfer level, and the program level which he also called the symbolic level, since it corresponds to the level of physical symbol systems he described with Simon in 1976. Newell noted that each of these levels has some essential attributes:

> "A level consists of a *medium* that is to be processed, *components* that provide primitive processing, *laws of composition* that permit components to be assembled into *systems*, and *laws of behavior* that determine how system behavior depends on the component behavior and the structure of the system. There are many variant instantiations of a given level, e.g., many programming systems and machine languages and many register-transfer systems."

Newell observes that levels can differ radically from each other, e.g. "the medium changes from electrons and magnetic domains at the device level, to current and voltage at the circuit level, to

bits at the logic level …, to symbolic expressions at the symbol level…" Newell also observes that all levels share four common features, which I quote:

> *Point 1*. Specification of a system at a level always determines completely a definite behavior for the system at that level (given initial and boundary conditions).
>
> *Point 2*. The behavior of the total system results from the local effects of each component of the system processing the medium at its inputs to produce its outputs.
>
> *Point 3*. The immense variety of behavior is obtained by system structure, i.e. by the variety of ways of assembling a small number of component types (though perhaps a large number of instances of each type).
>
> *Point 4*. The medium is realized by state-like properties of matter, which remain passive until changed by the components.

He discussed several other remarkable characteristics of computer systems levels, some of which are mentioned in the following pages. His review of these characteristics is interesting even 36 years later.

2.2.2 Newell's Surprising Proposal for the Knowledge Level

Newell (1982) proposes the existence of a knowledge level above the symbolic level. However, the proposed knowledge level has profoundly different characteristics (which he called surprises) from the symbolic level and other computer system levels. He states that the common features of computer system levels (points 1 through 4 in section 2.2.1 above) are not satisfied by his proposed knowledge level.

Newell specified many important theoretical and practical limitations for the knowledge level: He defined an agent at the knowledge level as a closed entity, with a very limited, fixed structure. No structural relationships are defined within an agent. There is no specification of processing mechanisms at the knowledge level, only a principle of rationality. Knowledge representation is not specified at the knowledge level. Newell explicitly precludes significant structure in the agent at the knowledge level. Any such structure must be represented at the symbolic level.

Newell proposed knowledge should be defined only indirectly as whatever can help predict the behavior of closed agents. There is no direct support in Newell's definition of the knowledge level for knowledge not directly involving behavior, e.g., knowledge of facts about the world, knowledge about knowledge, etc. There is no direct support at the knowledge level for representation in agents of other aspects of minds, such as ideas, theories, emotions, perceptions, etc. There is no direct support for representing agents as being able to communicate via languages. Newell specified that all representation must occur at the symbolic level.

2.2.3 Error and Limitations of Newell's Proposal for the Knowledge Level

To err is human. To identify important errors and correct them is a scientific duty. Science has progressed through history when misconceptions of great scientists have been identified and corrected. Thus while I have great respect and admiration for Newell's research, it is my scientific duty to say his proposal of a surprising nature for the knowledge level was a mistake, and his argument for its necessity was invalid. It is also a scientific duty to show how the important theoretical and practical limitations of his proposal can be avoided.

Newell (1982, pp.110-111) proposes an agent's knowledge is all the possible consequences of what the agent knows. This logical closure could be infinite. He acknowledges his proposal has little standing, citing counter-statements by Hintikka (1962) and Moore (1980). Newell considers the objection that a finite agent cannot have unbounded knowledge. He argues (pp.107-108) that the objection fails and that his definition of the knowledge level is necessary because an observer cannot set a bound on the set of propositions an agent knows.

Newell's argument is erroneous: The real knowledge of the world is captured in finite structures, i.e. brains, AI systems, and media. We frequently do not realize consequences of what we know, and forget things we used to know. It takes time to realize consequences of what we know. It is not necessary for us to be able to state a bound on the knowledge a human brain can hold, to conclude our knowledge is finite based on the facts that our brains are finite and our lifetimes are finite. By the same reasoning it is not necessary to state a bound on the amount of knowledge any finite computer system can hold to conclude that its knowledge will be finite.

Newell's proposed definition of the knowledge level is not just surprising it is *problematic*, for many reasons: The necessity of his proposal is justified by an erroneous argument. To the extent that the proposed knowledge level can exist, it may exist only as an ideal 'point at infinity', as a behaviorist ideal for agents with infinite knowledge. An infinite knowledge level only "approximates" what human intelligence can actually know, and the knowledge level can only be approximated by finite, physical symbolic systems: It cannot be realized by finite systems at the symbolic level. The definition of the knowledge level does not provide any specific guidance or direction for AI research and development. Newell says only that these efforts must continue to occur at the symbolic level. There is no opportunity for research and systems development within the knowledge level. Again, Newell explicitly precludes significant structure in an agent at the knowledge level.

A statement by Newell may indicate some reviewers disagreed with his proposal for the knowledge level. At the end of his 1982 paper he acknowledged extensive comments on an early draft by Jon Bentley, Daniel Bobrow, Hsiang-Tsung Kung, John McCarthy, John McDermott, Gregory Harris, Zenon Pylyshyn, Michael Rychener, and Herbert Simon. Newell wrote "They all tried in their several (not necessarily compatible) ways to keep me from error."

I would expect McCarthy probably gave objections along the lines I have given above, but I do not know. I cannot guess other reviewers' comments. Perhaps Newell decided it would be a contribution to science to advance a controversial proposal, lacking a better alternative.

It would be preferable to have a level above the symbolic level that has relationships to the symbolic level similar to the relationships the symbolic level has to lower computer system levels. It would be preferable for the level above the symbolic level to provide functional differences and advantages for representing and achieving human-level AI, not addressed directly at the symbolic level. In section 3 of this paper I will give a proposal for such a level.

2.3 Newell's 1990 Book *'Unified Theories of Cognition'*

2.3.1 Discussion of the Knowledge Level

Newell (1990) continues to advocate his 1982 definition of the knowledge level. He further discusses the issue of potentially infinite knowledge, writing (p.80):

> "It can be said that knowledge systems inherently can't be realized by symbol systems, because no finite device can realize an infinite set. This is true, but so are the inherent reasons why the other levels cannot be faithfully realized – why there cannot be computers with zero failure rates, or why settling time of digital circuits cannot be guaranteed to be finite. The pressing questions of approximation are how to obtain good approximations in particular cases and in practice. To this there can be no simple or single answer. The entire field of artificial intelligence is, in a sense, devoted to discovering the symbol-level mechanisms that permit a close approximation to the knowledge level."

However, there are qualitative differences between the nature of approximations for computers and digital circuits, and the question of how to approximate infinite knowledge. Engineers can successfully use physical materials and processes to approximate designs for digital computers well enough to achieve very low failure rates and very high performance in performing symbolic computation.

Newell (1990) apparently does not define what it would mean for a finite symbolic system to achieve a "close approximation" to an infinite knowledge level. The knowledge level remains a closed, infinite ideal, with the faults and limitations discussed in sections 2.2.2 and 2.2.3 above. He distinguishes between a system using all its knowledge or only having partial use of its knowledge. He then writes "A system is *intelligent* to the degree that it approximates a knowledge-level system." (p.90) He defines "perfect intelligence" as a system using all the knowledge it has to achieve its goals, even if a system has very little knowledge.

He acknowledges this definition has a counterintuitive property: A thermocouple maintaining room temperature can have perfect intelligence, while humans have imperfect intelligence. "To obtain systems with perfect intelligence, one must shrink down to very limited systems in terms of their knowledge and goals…" (p.94)

2.3.2 Discussion of 'Bands of Action'

In Chapter 3, Newell (1990) discussed different time scales at which human action can occur, due to processing within human brains, and communication between humans. He grouped these time scales into the following "bands of action":

Biological Band
- Organelle (10^{-4} secs)
- Neuron (10^{-3} secs)
- Neural Circuit (10^{-2} secs)

Cognitive Band
- Deliberate Act (10^{-1} secs)
- Cognitive Operations (1 sec)
- Unit Task (10 sec)

Rational Band
- Tasks ranging from minutes to hours

Social Band
- Actions ranging from days to months

These bands of action correspond to real, physical events that happen as a result of processing within finite human brains. Newell's (1990) discussion of these bands of action was quite different and separate from the closed theoretical ideal of a potentially infinite knowledge level that he proposed in 1982 and continued to advocate in 1990. I give a discussion of the bands of action in (Jackson, 2018b), and mention them later in this paper (section 3.7).

2.3.3 Suggestions for Unified Theories of Cognition

Newell (1990) advocated developing 'unified theories of cognition' which "gain their power by positing a single system of mechanisms that operate together to produce the full range of human cognition". He gave an initial list of areas to be covered by a unified theory (p.15):

> Problem solving, decision making, routine action
> Memory, learning, skill
> Perception, motor behavior
> Language
> Motivation, emotion
> Imagining, dreaming, daydreaming...

Newell made it clear this list of areas for a unified theory was incomplete by using the ellipsis and writing (on p.17) "any such theory will be inadequate and so it will change and grow". Thus, page 19 listed additional topics, including:

> Use symbols and abstractions
> Use language, both natural and artificial
> Learn from the environment and from experience
> Be self-aware and have a sense of self

Newell also made clear that unified theories of cognition should be reflected in working computer systems, writing "Calculations and simulations count".

While I disagree with Newell's definition of the knowledge level, I agree we should strive for unified theories of cognition. His advocacy of unified theories was an important step toward human-level artificial intelligence. The discussion in the following pages of this paper is consistent with Newell's goals for unified theories of cognition.

3. Future Research Directions

3.1 Identifying A New 'Level of Thought': The Intelligence Level

If we stipulate Newell's definition of the knowledge level was problematic, how should it be replaced? What if anything can be defined above the symbolic level, and shown theoretically to exist above the symbolic level, that would provide functional differences and advantages for representing and achieving human-level AI, which are not already addressed at the symbolic level? Is there an 'intermediate level of abstraction' between the symbolic level and Newell's

problematic ideal of potentially infinite knowledge, which is useful in describing and supporting work on human-level AI?

As a first step, let us reconsider what is provided at the symbolic level: It is defined as the level at which symbolic processing occurs. Newell and Simon's 1976 paper essentially defined the symbolic level as the level at which programming languages exist, and in which programs are defined and executed. It is the level of universal computation.

However as noted above there is a normal limitation for systems at the symbolic level which Newell and Simon (1976) did not discuss. Typically, a computer program has no understanding of what it is doing or what its symbols mean in the external, real world. Human intelligence has been needed to insure that computer programs correctly perform the processes in the real world which humans intend for programs to accomplish. Typically, computer programs cannot debug themselves to perform as people intend.

Because such understanding is not intrinsic to the symbolic level, it follows that human-level intelligence operates at a system level above the symbolic level, to create useful programs at the symbolic level. If human-level artificial intelligence is achieved, it could also operate at this system level above the symbolic program level, creating useful programs at the symbolic level.

The potential existence of human-level AI at a level above the symbolic level thus follows from the characteristic that "Each computer system level is a specialization of the class of systems capable of being described at the next lower level." (Newell, 1982) Another way of stating Newell and Simon's Physical Symbol System Hypothesis is to say that human-level artificial intelligence is a specialization of physical symbol systems: some physical symbol systems can have human-level intelligence, hypothetically, even though not all physical symbol systems do.

For concision, I will call this new system level *'the intelligence level'*, with the understanding that it refers to human-level intelligence and hypothetically human-level AI. This name will distinguish the new level from Newell's proposal for the 'knowledge level'. The intelligence level may be considered as a 'level of thought' above the symbolic level.

To summarize: There is at least one typical limitation at the symbolic level which Newell did not discuss in developing his proposal for an ideal, potentially infinite knowledge level. Human intelligence overcomes this typical limitation and operates at a level above the symbolic level. Newell's problematic definition for the knowledge level is not necessary. A computer systems level that may actually exist above the symbolic level is human-level artificial intelligence, i.e. the intelligence level. The following pages will discuss how the intelligence level may be a useful level of abstraction for work on human-level AI.

3.2 Abilities Required for Systems at the Level of Human Intelligence

To proceed further, it will be helpful to briefly discuss what the abilities of human intelligence imply for systems at the intelligence level.

3.2.1 What Would It Mean for Computer Programs to Understand What Their Symbols Mean?

We can imagine that if a human-level AI were asked to write an accounting program for a corporation, it would write a program in some programming language (e.g. Java) with symbols (e.g. object names, variables, constants…) that would represent employees, products, vendors,

suppliers, sales, purchases, taxes, financial accounts, etc. The human-level AI could explain in English to people what these symbols meant, and explain how and why the program worked as it did. The human-level AI could understand explanations people would give in English about how the accounting program should treat employees, vendors, suppliers, etc. The human-level AI could find bugs in the accounting program or be told the program had a bug, and find a way to change the program to remove the bug.

In doing these things, the human-level AI would operate at the intelligence level, above the symbolic systems level of an ordinary computer program that cannot do these things. This is the classic 'symbol grounding' problem in a very real context: ordinary computer programs do not understand how their symbols are grounded, relative to things that people believe exist in reality.

In principle, the human-level AI could also inspect and understand the computer program that defined the human-level AI itself, and find ways to improve its own performance. Such self-programming would also occur at the intelligence level, above the symbolic systems level of ordinary computer programs that cannot modify and improve themselves. People are also able to change and improve their ways of thinking and doing things, at least to some extent.

This does not mean that a human-level AI must have complete self-knowledge and be able to explain what all its symbols and processes mean, without infinite recursion. Whether that is possible in principle is a theoretical question for another day.

3.2.2 What Other Abilities Should Systems Have at the Intelligence Level?

There are several other, related abilities that human-level intelligence has, which symbolic systems in general do not have. These abilities will be necessary for a system to achieve human-level artificial intelligence, and are appropriate to include as abilities of systems at the intelligence level. Following is an initial, brief discussion. In general, these features are called *'higher-level mentalities'* in (Jackson, 2014) to characterize objectives for research on human-level AI.

If human-level AI is achieved, these abilities would be implemented by computer programs that would be different from the much wider set of programs that do not implement these abilities. Hence these programs would be examples of systems at the intelligence level, above the symbolic systems level of an ordinary computer program that cannot do these things.

3.2.2.1 Natural Language

Humans need a natural language like English to develop and share an understanding of the world in virtually all its aspects, to understand humans who use natural language, and to explain its thoughts to humans. Attempts to build systems that process 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.

3.2.2.2 Higher-Level Learning

The ability to learn is another key feature of systems at the intelligence level. There has been much previous research on machine learning which qualifies as forays into the intelligence level (Valiant, 2013). However, there is still much work needed to achieve the 'higher-level learning' shown by human intelligence. I use this term to refer collectively to forms of learning such as

learning by creating explanations and testing predictions about new domains based on analogies and metaphors with previously known domains, reasoning about ways to debug and improve behaviors and methods, learning and invention of natural languages and language games, learning or inventing new representations, and in general, self-development of new ways of thinking.

3.2.2.3 Meta-Cognition

The discussions above can be extended to include meta-cognition in general, i.e. "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, for example:

- Reasoning about reasoning.
- Reasoning about learning.
- Learning how to reason.
- Learning how to learn.
- …

3.2.2.4 Imagination

The ability to imagine is another key feature of systems at the intelligence level. 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. Imagination is another hallmark of human intelligence, which a human-level AI should demonstrate.

3.2.2.5 Creativity and Originality

A key feature of human intelligence is the ability to create original concepts. The test of originality should be whether the system can create (or discover, or accomplish) something for itself it was not taught directly -- more strongly, in principle and ideally in actuality, can it create something no one has created before, to our knowledge? This is Boden's (2004) distinction of (personal, psychological) P-creativity vs. (historical) H-creativity.

3.2.2.6 Spatial-Temporal Reasoning and Visualization

Related to imagination, people have the ability to visualize situations in three-dimensional space and reason about how these situations might change, e.g. by visualizing motions of objects. This ability is very important for planning physical actions, and also important for imagination, and for understanding natural language.

3.2.2.7 Self-Awareness – Artificial Consciousness

A system lacking some awareness of its own existence, or some awareness of what it is doing, would probably not be considered to have human-level intelligence by most people.

To provide such awareness it is not necessary to solve all the mysteries of human consciousness, or for AI systems to achieve the subjective experience humans have of consciousness. The 'Hard Problem' of consciousness (Chalmers, 1995) is the problem of explaining the first-person, subjective experience of consciousness. This is a difficult, perhaps scientifically unsolvable problem because science relies on repeatable experiments and second-and-third-person explanations. Evidently there is not a philosophical or scientific consensus for the Hard Problem.

Rather than trying to solve the Hard Problem for systems at the intelligence level, I advocate adapting Aleksander and Morton's (2007) "axioms of being conscious" for research on 'artificial consciousness'. For a system to have artificial consciousness it should demonstrate:

- Observation of an external environment.
- Observation of itself in relation to the external environment.
- Observation of internal thoughts.
- Observation of time: of the present, the past, and potential futures.
- Observation of hypothetical or imaginative thoughts.

To observe these things, a human-level AI should have representations of them, and support processing such representations. The nature of such representations will be briefly discussed below. Artificial consciousness and consciousness in general are discussed in more detail in (Jackson, 2014), viz. §2.1.2.8, §2.3.4, §3.7.6, §4.2.7, §6.3.6. [1]

3.2.2.8 Emotions and Emotional Intelligence

A human-level AI will need some level of social understanding to interact with humans. It will need some understanding of cultural conventions, etiquette, politeness, etc. It will need some understanding of emotions humans feel, and it may even have some emotions of its own, though we will need to be careful about this. One of the values of human-level artificial intelligence is likely to be its objectivity, and freedom from being affected by some emotions.

However, this paper will not discuss how emotions and emotional intelligence could be represented in a human-level AI, other than to note that natural language can be used to describe emotions.

3.2.2.9 Ethical Values and Ethical Intelligence

A human-level AI will need an understanding of ethical values and ethical rules and principles to interact with humans, and to support "beneficial AI" – AI that is beneficial to humanity and to life in general. (Bringsjord, Arkoudas, & Bello, 2006; Tegmark, 2017) This is a topic that has become increasingly important, as people have considered the potential good and bad consequences AI might have for humanity (Jackson, 2018a).

[1] Throughout this paper the § notation is used to refer to chapters and sections in (Jackson, 2014). For example, §2.1.2.8 refers to Chapter 2, section 1.2.8. These references can be directly accessed from the Table of Contents in (Jackson, 2014) and via hyperlinks within the thesis.

3.2.2.10 Virtues

There may be no reason in principle why we would not want a human-level artificial intelligence to possess a virtue such as wisdom, kindness, or courage, if the situation merited this. However, there may be no simple or valid way to define these virtues well enough to specify them as behaviors of an artificial system. We may hope these virtues might result from behaviors of systems that have emotional and ethical intelligence. This is another topic for future consideration.

3.3 Requirements for Representation and Processing at the Intelligence Level

Having discussed abilities required for systems at the intelligence level, it is appropriate to give some further discussion about how these abilities can be achieved, and what would be the nature of representation and processing required at this level.

Newell (1982) explicitly precluded significant structure in the agent at the knowledge level. In contrast, the design of agents at the intelligence level is open for research and development. Our creation of human-level artificial intelligence is a work in progress. We do not know enough yet to define everything at the intelligence level. Yet we can discuss initial requirements for representations and processing.

3.3.1 What Kinds of Representations Must Exist at the Intelligence Level?

Considering the abilities systems must have at the intelligence level, discussed in section 3.2 above, we can list the following things which systems at the intelligence level must be able to represent:

- An intelligent system must be able to represent meanings of symbols. More generally, a system must be able to represent meanings of natural language words and expressions.
- An intelligent system must be able to represent concepts that describe procedures, e.g. corresponding to computer programs, or corresponding to procedures people may describe in natural language. I call these ‘executable concepts’, or ‘xconcepts’.
- To understand whatever exists in the world, an intelligent system must be able to represent whatever can exist, and to represent different modes of existence, e.g. objective, subjective, intersubjective, hypothetical, or fictional existence…
- To support higher-level learning, an intelligent system must be able to represent explanations and predictions about domains of knowledge, analogies and metaphors between domains.
- An intelligent system must be able to represent states and behaviors of procedures, to debug and improve procedures.
- An intelligent system must be able to represent languages, to learn and invent languages.
- An intelligent system must be able to represent representations, in order to reason about representations and invent new representations.

- An intelligent system must be able to represent ideas and thought processes, in order to develop new ways of thinking, and to support meta-cognition.
- An intelligent system must be able to represent that it does not know something, or does not know how to do something.
- An intelligent system must have the ability to represent spatial-temporal situations, to represent actions and events within such situations, and to represent alternative ways such situations may evolve.
- An intelligent system must have the ability to represent its external environment, and its relationships to the external environment.
- An intelligent system must have the ability to represent its internal thoughts, and refer to its internal thoughts within other thoughts. It must be able to think about its thinking.
- An intelligent system must have the ability to represent time: the present, the past and potential futures.
- An intelligent system must have the ability to represent situations and contexts, including hypothetical and imaginative situations or contexts.
- An intelligent system must have the ability to represent emotions people may have, and perhaps represent emotions it may have.
- An intelligent system must have the ability to represent ethical values, rules, and principles.
- Ideally, an intelligent system should have the ability to represent virtues such as wisdom, kindness, or courage.

This is a daunting list, but it can be simplified by realizing that many items in the list can be represented using natural language expressions, or involve representation of semantics, contexts, spatial and temporal relations, etc., which may be referenced by natural language expressions.

To achieve human-level artificial intelligence, an AI system must somehow represent the full range of human thoughts. No existing formal language can do this. A natural language like English already has the ability to do this, arguably as well as any artificial, formal language could.

So, much of what needs to be represented at the intelligence level may be addressed if we can develop an adequate framework for representing the general semantics and syntax of natural language. I will suggest how this may be done, beginning in section 3.4 below.

3.3.2 What Kinds of Processing Must Exist at the Intelligence Level?

Since all the forms of representation listed above would be symbolic representations in a physical symbol system, a simple answer is to say that the symbolic processing needed to create, modify, and use these symbolic representations must exist at the intelligence level. This symbolic processing may be organized into a cognitive cycle, which will be discussed below. Each of the various representations may require processing specific to the representation.

It should again be expressly noted that these symbolic processes and representations can (at least hypothetically) be implemented at the level of physical symbol systems. It is not being

claimed otherwise. All that is being claimed is that these symbolic processes and representations which may achieve human-level AI are a special case of the much wider set of processes and representations at the physical symbol systems level which in general do not achieve human-level AI. Hence the symbolic processes and representations to achieve human-level AI characterize the intelligence level, as a specialization of the physical symbol systems level.

3.4 Design of an Architecture and Representation Language for Human-Level AI

Having discussed requirements for representation and processing at the intelligence level, I will next discuss how these requirements could be addressed by an architectural framework for human-level AI and for representing meaning in natural language. This will be a brief introduction to the 'TalaMind' architecture I have proposed for research toward human-level AI (Jackson, 2014, §1.5).

In the next section, I will describe the design, processing, and output of a prototype demonstration system for the TalaMind architecture and representation language – some information about the prototype is also given in the next few paragraphs.

I conjecture that a natural language like English is the best general representation language for systems at the intelligence level. The TalaMind hypotheses (§1.4) advocate developing a human-level AI using a language of thought (called Tala) based on the unconstrained syntax of a natural language; designing the system as a collection of 'executable concepts' that can create and modify concepts, expressed in the language of thought, to behave intelligently in an environment; and using methods from cognitive linguistics such as mental spaces and conceptual blends for representing semantics, metaphors, and analogies (Fauconnier & Turner, 2002), §3.6.7.8, §3.6.7.9.

The theoretical basis for the Tala conceptual language is discussed in Chapter 3 of the TalaMind thesis; §3.3 argues it is theoretically possible to use the syntax of a natural language in a conceptual language and to reason directly with natural language syntax and semantics. With the TalaMind approach, it is not necessary (though it is possible) to translate Tala natural language expressions to and from formal languages such as predicate calculus, conceptual graphs, etc., for programs to reason about meanings of natural language.

This is illustrated in the TalaMind prototype demonstration system, which includes pattern-matching logic for Tala expressions to support inference with natural language syntax. Tala uses natural language syntax for representing conceptual expressions, and generalizes production rules as executable concepts expressed in Tala. In the TalaMind prototype these are supported with cognitive cycles for pattern-matching of Tala expressions.

The TalaMind architecture has three levels of conceptual representation and processing, called the linguistic, archetype, and associative levels, adapted from Gärdenfors' (1995) paper on levels of inductive inference. These might be considered as sub-levels within the intelligence level.

At the linguistic level, the architecture includes the Tala language, and also a 'conceptual framework' for managing concepts expressed in Tala, and conceptual processes that operate on concepts in the conceptual framework to produce intelligent behaviors and new concepts. The archetype level is where cognitive categories are represented using methods such as conceptual spaces, image schemas, radial categories, etc. The TalaMind approach follows a consensus

perspective of work on cognitive linguistics and cognitive semantics (Evans & Green, 2006). The associative level would typically interface with a real-world environment and support connectionism, Bayesian processing, etc. In general, the thesis does not prescribe specific research choices at the archetype and associative levels.

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

A Tala expression is a multi-level list structure representing the dependency parse-tree (syntax) of a natural language expression. 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 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: Tala expressions can include pointers to word senses and referents. A further discussion of natural language in the TalaMind approach is given by (Jackson, 2018c). Further discussions about meaning and representation are given in §2.2.2, §3.6, §4.2.8, §5.4.

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) which Newell (1990, p.155) criticized. This broader, generalized sense corresponds more to a paper by Doyle (1983), and refers to a multiagent system using a language of thought for communication (§2.3.3.2.1).

In the TalaMind 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) within a Tala agent, using Tala as an interlingua. Viz. (Baars & Gage, 2007) and §3.6.7.13, §4.2.7, §5.4.4, §6.3.6.

The TalaMind architecture is actually a broad class of architectures, open to further design choices at each level (§1.5, §2.2.2). Comparisons of TalaMind to the Common Model of Cognition are given by (Jackson, 2018c) and (Jackson, 2017).

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

The Tala language responds to McCarthy's (1955) proposal for a formal language that corresponds to English (§1.1) though not in the way McCarthy sought. Tala enables a TalaMind system to formulate statements about its progress in solving problems. Short English expressions have short correspondents in Tala, a property McCarthy sought for a formal language in 1955. Tala can represent unconstrained, complex English sentences, involving self-reference and conjecture. Thesis chapter 4 discusses theoretical objections, including McCarthy's arguments in 2008 that a language of thought should be based on mathematical logic instead of natural language (§4.2.5) and Searle's Chinese Room argument (§4.2.4).

Of course, much further work is needed to achieve human-level AI via the TalaMind approach (§7.7).

3.5 A Prototype Demonstration System for the Intelligence Level

To illustrate further how human-level AI may eventually be achieved by developing systems at the intelligence level, this section briefly describes the design, processing, and output of the TalaMind prototype demonstration system (Jackson, 2014).

3.5.1 Nature and Design of the Prototype Demonstration System

The demonstration system is a functional prototype in which two Tala agents, named Ben and Leo, interact in a simulated environment.

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.

The demonstration system includes a prototype design for a conceptual framework and conceptual processes at the linguistic level of a TalaMind architecture, for each Tala agent. The conceptual framework includes prototype representations of perceived reality, subagents, 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.

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. It was not possible within the timeframe and resources available for the thesis to implement the archetype or associative levels of a TalaMind architecture. Abstracting out these levels facilitated creation of the prototype.

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. In each case, the stories

that are simulated are essentially pre-defined: what happens depends on the initial goals, knowledge, and executable concepts that Ben and Leo have within their conceptual frameworks. Thesis chapter 6, section 4 discusses how these story simulations illustrate the potential of the TalaMind approach to support higher-level mentalities in human-level AI.

I wrote the TalaMind prototype demonstration software in JScheme and Java.

3.5.2 The Discovery of Bread Story 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 next 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.[2] A prototype routine called FlatEnglish converts Tala expressions into English text output displayed by the simulation, creating some typographical errors in the output.

3.6 Relation of the Intelligence Level to Previous Research

As previously noted, research efforts on machine learning count as forays into the intelligence level even though such research has not achieved human-level AI. Learning is a hallmark of intelligence and the successes in machine learning are steps toward human-level AI.

However, significant research is still needed to realize 'higher-level learning' as described in section 3.2.2.2 above. Such learning is important for human-level AI. The TalaMind approach suggests ways to implement higher-level learning.

Likewise, research efforts on systems which use ontologies, and/or support understanding natural language even to a limited degree, count as forays into the intelligence level. The same holds for research on representing commonsense knowledge, or research on developing cognitive architectures. These research efforts all count as forays into the intelligence level.

---

[2] The story simplifies the process for making bread, and omits steps of threshing and winnowing grain, describing just a single step "pounding grain". The use of beer foam to leaven bread does have an historical basis: Pliny the Elder wrote that the people of Gaul and Spain used the foam from beer to leaven bread and "hence it is that the bread in those countries is lighter than that made elsewhere" (Bostock & Riley, 1856, IV, Book XVIII, p. 26).

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!

Newell's (1990) advocacy for developing unified theories of cognition counts as advocating research on the intelligence level, even though he did not recognize the existence of the intelligence level as a potential computer systems level above the physical symbol systems level. Although he listed language as a subject for unified theories of cognition, he intentionally did not delve into language, writing (p.16) that it should "be approached later rather than sooner".

Newell's (1990) discussion of 'bands of action' for human cognition also counts as advocating research on the intelligence level, even though the bands of action were described relative to processing in the human brain rather than computers, and he somewhat understated the social band of action's importance for cognition (Jackson, 2018b).

A general observation: Much previous AI research has treated natural language as an application to be supported using simpler formal languages for internal data and rules of inference. An argument can be made that such approaches are unlikely to succeed in achieving human-level AI, because natural language representation and processing is a core functionality of human-level intelligence, arguably needed in some fashion for internal representation of thought (Jackson, 2018c).

4. Conclusion

This paper has shown why and how Newell's 1982 definition of a 'knowledge level' should and can be replaced by a definition of an 'intelligence level' corresponding to human-level intelligence, and potentially human-level artificial intelligence. The TalaMind approach (Jackson, 2014) has been described as a direction for research and development at the intelligence level.

The challenge for work within the intelligence level remains to achieve Newell and Simon's vision, and the vision of McCarthy, Minsky, Rochester, and Shannon who conjectured "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

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

Acknowledgements

I thank Patrick Langley for editorial guidance and theoretical questions about this paper, which was originally titled *Thoughts on Levels of Thought*. Three anonymous reviewers gave questions and criticisms which were helpful in revising the paper.

Appendix I. Relation of the Intelligence Level to Lower Computer System Levels

Since the intelligence level has been identified in this paper as a hypothetical new computer system level, for completeness some discussion should be given about how the intelligence level relates to lower computer system levels, according to Newell's (1982) discussion of these levels.

Newell (1982) wrote that each computer system level involves processing a medium. So, a natural question is "What is the medium processed at the intelligence level?"

I would describe the medium of the intelligence level as "concepts", defined very broadly in (Jackson, 2014, §1.5) as "ranging from immediate thoughts and percepts to long term memory, including concepts representing definitions of words, knowledge about domains of discourse, memories of past events, etc." Other equally good words would be "thoughts" or "ideas", if they are understood to have the same broad range of reference. For now, I will simply use "concepts", with the condition that concepts / thoughts / ideas are represented symbolically, e.g. as proposed in the TalaMind approach.

The intelligence level has the four common features Newell (1982) observed for lower computer system levels, quoted above in section 2.2.1 of this paper. The TalaMind approach adds some features:

1. Behavior can also be affected by the concepts a system has.
2. Behavior can also result from processing executable concepts.
3. Variety of behavior can also be affected by the concepts that a system has.
4. Concepts can also be changed when executable concepts are processed.

The intelligence level has the same relationships to the symbol level as the symbol level has to lower levels, which Newell (1982, p.95) discussed in two paragraphs beginning "Each level is defined in two ways". Page limits for this paper prevent going into further details.

Newell (1982) wrote that each computer system level has laws of behavior. Behaviors at the intelligence level are a topic for research in developing human-level AI. Newell's principle of rationality is a starting point, yet intelligent behaviors also involve learning, play, imagining, understanding, communicating, emotions... So, I will not try to identify all the laws of behavior involved in processing at the intelligence level, or discuss them further in this appendix.

A general capability can be provided by a computer systems level even though it specializes the next lower computer systems level. Thus the symbolic processing level provides universal computation even though it specializes the hardware level below. The intelligence level will hypothetically provide the general capabilities of human-level intelligence, while preserving the ability to understand programs and support universal computation, by specializing computer programs to those which can support human-level AI.

References

Aleksander, I., & Morton, H. (2007). Depictive architectures for synthetic phenomenology. In A. Chella & R. Manzotti (Eds.), *Artificial consciousness*. Exeter, UK: Imprint Academic.

Baars, B. J., & Gage, N. M. (2007). *Cognition, brain, and consciousness – Introduction to cognitive neuroscience*. Amsterdam: Elsevier.

Boden, M. A. (2004). *The creative mind – Myths and mechanisms*. London: Routledge.

Bostock, J., & Riley, H. T. (1856). *The natural history of Pliny*. London: H. G. Bohn.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, *6*, 38−44.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 200−219.

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).

Evans, V., & Green, M. (2006). *Cognitive linguistics – An introduction*. London: Lawrence Erlbaum Associates.

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. Amsterdam: Elsevier.

Harari, Y. N. (2015). *Sapiens: A brief history of humankind*. New York: HarperCollins.

Hintikka, J. (1962). *Knowledge and belief*. New York: Cornell University Press.

Jackson, P. C. (2014). *Toward human-level artificial intelligence – Representation and computation of meaning in natural language*. Doctoral dissertation, Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands.

Jackson, P. C. (2017). Toward human-level models of minds. *AAAI Fall Symposium Series Technical Reports*, FS-17-05, 371−375.

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

Jackson, P. C. (2018b). *Thoughts on bands of action*. Unpublished manuscript, TalaMind LLC, Troy, Michigan.

Jackson, P. C. (2018c). *Natural language in the common model of cognition*. Unpublished manuscript, TalaMind LLC, Troy, Michigan.

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

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence.* Reprinted in R. Chrisley & S. Begeer (Eds.) (2000) A*rtificial intelligence: Critical concepts in cognitive science, 2*, 44−53. London: Routledge.

Minsky, M. L. (1986). *The society of mind*. New York: Simon & Schuster.

Moore, R. C. (1980). *Reasoning about knowledge and action* (Technical Note 191). SRI International, Menlo Park, CA.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*, 87−127.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Upper Saddle River, NJ: Prentice-Hall.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*, 113−126.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. New York: Alfred A. Knopf.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433−460.

Valiant, L. G. (2013). *Probably approximately correct – Nature's algorithms for learning and prospering in a complex world*. New York: Basic Books.