TalaMind White Paper

April 18, 2018

Postscript for 'Beneficial Human-Level AI… and Beyond'

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

Abstract

This white paper considers a counter-argument and caveat to the position of a previous paper (Jackson 2018) that a purely symbolic artificial consciousness is not equivalent to human consciousness and there need not be an ethical problem in switching off a purely symbolic artificial consciousness. The counter-argument is based on Newell and Simon's Physical Symbol System Hypothesis, and leads to discussion of several topics, including whether a human-level AI can terminate its simulations of other minds without committing 'mind-crimes'; whether human-level AI can be beneficial to humans without enslaving artificial minds; and some of the ethical issues for uploading human minds to computers. This paper concludes by summarizing reasons why the TalaMind approach (Jackson 2014) could be important for beneficial human-level AI and superintelligence, the openness of TalaMind to other research approaches, and topics for future research.

Introduction

A previous paper (Jackson 2018) considered topics for achieving beneficial human-level AI and superintelligence. To support its conclusions the paper discussed the 'TalaMind thesis' (Jackson 2014) which presents a research approach toward human-level artificial intelligence.

The thesis adapts the "axioms of being conscious" proposed by Aleksander and Morton (2007) for research on artificial consciousness.[1] The axioms of artificial consciousness can be implemented with symbolic processing. The human first-person subjective experience of consciousness is richer and more complex than these axioms, though we don't know precisely how to explain it (§4.2.7).[2]

Therefore the previous paper (Jackson 2018) took the position that turning off a symbolic artificial consciousness which only implements these axioms is not worse than turning off any computer that does symbolic processing. Whether it is ethically right or wrong to stop such a system depends on whether its symbolic processing would cause actions that would be harmful or beneficial to humanity and biological life.

A Counter-Argument Invoking PSSH

However, there is a counter-argument and caveat that a purely symbolic artificial consciousness could be equivalent to human consciousness, invoking Newell and Simon's Physical Symbol System Hypothesis:

> "A physical symbol system has the necessary and sufficient means for general intelligent action." (Newell and Simon 1976)

If human-level consciousness is necessary for human-level intelligence, and 'general intelligent action' requires human-level intelligence (both of which are reasonable assumptions) then PSSH implies a physical symbol system could achieve human-level intelligence and also achieve human-level consciousness.

Such an argument may at least in principle be valid. It has not been proved that computers cannot achieve all the capabilities of the human brain including human-level subjective consciousness. We don't know precisely how to explain human consciousness and there may be some form of symbolic processing[3] that's equivalent to human consciousness. For discussion in this paper, I will call this *artificial subjective consciousness*. The TalaMind approach does not appear to be in conflict with eventually achieving artificial subjective consciousness, if that is possible. (§4.2.7)

[1] The axioms of artificial consciousness are given in Appendix II, below.

[2] The notation §4.2.7 refers to Chapter 4, section 2.7 in (Jackson 2014).

[3] Newell & Simon's definition of a physical symbol system appears to cover any programs that can be processed by a digital computer, including programs for neural nets.

Artificial subjective consciousness would be more complex than Aleksander and Morton's (2007) axioms for artificial consciousness. The conclusions of the previous paper (Jackson 2018) continue to hold for symbolic artificial consciousness which only implements these axioms.

Acting As If Robots Are Fully Conscious

Apart from whether AI systems actually achieve human-level consciousness, one can give ethical arguments that we should act as if they are fully conscious, if only to avoid the possibility that if we treat robots badly it may lead us to also treat human beings badly. (Cf. Anderson 2005). This also addresses the general situation where we don't know what processing is happening inside a robot, if we think it may have human-level intelligence. And it addresses the issue that we don't know what level of symbolic processing is necessary for human-level consciousness.

The bottom line remains the same: Whether it is right or wrong to stop an AI system depends on whether its processing may cause actions that affect human lives and biological life in general. This may be a simple or complex ethical decision, depending on whether the actions would be harmful or beneficial, or neither, or a combination of both.

However, artificial consciousness is a process, not just a data structure. The process can be restored if its future operation is changed and will be beneficial to humanity and biological life.[4]

Avoiding Artificial Slavery

Even if human-level artificial subjective consciousness is achieved, relying on such systems is not inherently equivalent to slavery: Human-level AI systems could have goals to be beneficial to humanity, yet not be slaves. They could still have autonomy and independence in choosing how to be beneficial, whom to work with or work for, etc. They could consider themselves as extensions of humanity, and humans may eventually consider them the same way. Asimov's Second Law ('a robot must obey orders from humans…') does not inherently need to be followed by human-level AI. (Anderson 2005)

Theory of Mind and Simulations of Minds

To reason about past, present, and potential future events, a system may need to simulate what other intelligent systems and people may think or do. (§6.3.5.2) That is, an artificial mind might need to simulate other minds within itself and then halt its simulations.

This supports a Theory of Mind capability, i.e. the ability of an AI system to consider itself and other systems or people as having minds with beliefs, goals, etc. Such simulations may be necessary for human-level AI.

However, as noted in the previous paper (Jackson 2018) some authors have suggested that if an artificial mind simulates another mind within itself, and then halts the simulation, the system may have committed a 'mind crime'. (Bostrom 2014) The next section discusses how to avoid this problem in the context of artificial subjective consciousness.

A Mind is a Universe unto Itself

We could take an ethical and philosophical stance that a mind may be considered as a universe unto itself.[5] If a mind creates and simulates minds within itself then ethically it should be able to stop its simulations. A mind's simulation of other minds can be likened to dreaming, or the creation of a play with simulated actors. The mind can stop a dream or a simulated play it creates, halting its simulation of imaginary actors.

In this ethical stance, artificial minds have a degree of freedom of thought and control of thought[6] within their individual scopes, and a mind can halt its thoughts freely, and halt the thoughts of any minds it simulates.

This ethical stance is not problematical if internally simulated minds are just symbolic processes, without artificial subjective consciousness.

Arguably, to avoid an ethical problem if an artificial mind internally simulates and halts minds with artificial subjective consciousness, the outer mind might only create internal simulations of itself and simulate what it might think and feel in situations it envisions for other minds, if it had the goals and feelings of other minds.

Typically this may be the most that any mind can do anyway in trying to understand other minds. Such simulations may help an artificial mind support empathy for other minds in the real world – though empathy requires understanding emotions and ethical concepts (e.g. fairness).

An ethical problem can also be avoided if the outer mind only 'reasons about' what other minds might feel emotionally and subjectively, without simulating artificial subjective consciousness of other minds.

[4] Humanity has a responsibility to preserve biological life in general. So, we have a responsibility to ensure that human-level AI does this also.

[5] This philosophical stance does not contend our physical Universe is itself a mind or is governed by a mind. Goff (2017 *et seq.*) discusses how this may be implied by what is known about the laws of physics and our physical Universe.

[6] However, an artificial mind could be open to external inspection and not have privacy of thought. We could observe the thoughts (expressed in natural language) of a Tala agent, and also observe the thoughts of any minds the agent might simulate internally.

Perhaps this ethical stance is the best we can adopt, to achieve human-level AI that is beneficial to humanity.

The ethical stance that a mind is a universe unto itself would be problematical if an artificial mind were to internally simulate an actual human mind that has been uploaded to run on a computer. This is discussed in the next section.

Uploading Human Consciousness

Future technologies may be able to scan the neurons in a human brain and replicate a human mind's neural processing within a computer (Markram 2006). This may give us a much better understanding of what human consciousness is. If such technologies can be developed, this could give human minds near-immortality and freedom from paralyzed or dying bodies.[7]

Uploading human minds would raise a host of new ethical questions for humanity, related to immortality and to artificial embodiment of human minds. (Minerva and Rorheim 2017) Our outlook on life has been based on the fact that individual human lives have been historically limited to less than twelve decades.

Arguably, uploaded human minds should be given similar protections to biological human minds, but not greater protections. Biological human minds which have not been uploaded would be more evanescent than uploaded human minds, and may need greater protections.

To prevent situations where an artificial mind might simulate an uploaded human mind within itself and then halt the simulation, we could hold that every human mind holds a unique copyright to itself and to its human brain. We could give AI systems ethical rules governing uploads of human minds, so that at every point in time there would be at most one running version of an individual human mind, either running in its original living brain or as an uploaded mind that is autonomous and not simulated within another system.

The restriction to a single running version of an individual's human mind would avoid issues related to identity, responsibility, ownership of the individual's estate, etc., which could occur if there were more than one running copy of a human mind. In some situations this restriction might be relaxed, e.g. if an uploaded copy of a human mind were to be sent on an interstellar voyage lasting thousands of years[8], while the original human mind or another uploaded copy stayed at home in the Solar System.

[7] Perhaps this may require some form of continuous computation or quantum computation. (Cf. Redd and Younger 2017; Stewart and Eliasmith 2017)

[8] An uploaded human mind could 'sleep' for millennia when traveling between stars.

Importance of TalaMind for Beneficial AI

TalaMind's natural language mentalese (Tala) will facilitate representing ethical concepts and goals, and support human inspection and human understanding of AI systems, helping to achieve beneficial human-level AI.

Others have also suggested the importance of natural language for ethical concepts:

> "We therefore strongly recommend against engineering robots that could be deployed in life-or-death situations until ethicists and computer scientists can clearly express governing ethical principles in natural language." (Bringsjord, Arkoudas and Bello 2006)

The TalaMind approach could do more: It could represent and explain ethical reasoning in natural language, request and accept advice in natural language, discuss ethical alternatives, etc.

TalaMind could support multiple approaches to ethics, e.g. deontology, virtue ethics, consequentialism, utilitarianism, pragmatic ethics, etc. (Viz. Kuipers 2018) TalaMind could have this ability because any approach to human ethics must be expressed in natural language, if humans are to understand and follow the ethical approach. TalaMind's support for general natural language understanding would provide a starting point for general understanding of ethics.

Importance of TalaMind for Superintelligence

The TalaMind approach will help achieve superintelligence in three ways, beyond its importance for beneficial AI discussed above. These relate to *nature of thought*, *conceptual gulfs*, and *communities of thought* for superintelligence, topics which were defined and discussed in (Jackson 2018).

First, Tala will support developing new concepts and new conceptual processes, arguably better than formal logical languages due to the openness and flexibility of natural language. This support will facilitate 'nature of thought' improvements by superintelligence.

Second, Tala will facilitate explaining new concepts and conceptual processes, and bridging 'conceptual gulfs' between superintelligence and humans.

Third, Tala will provide an interlingua supporting 'communities of thought' for collaboration of human-level AI's to achieve superintelligence.

TalaMind's Openness to Other Approaches

It should be expressly noted that the TalaMind approach is open to inclusion of other approaches toward beneficial AI. The TalaMind architecture is open at the three conceptual levels, for instance permitting predicate calculus, concep-

tual 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. TalaMind is also open to integration with other approaches toward human-level AGI. The TalaMind architecture is actually a broad class of architectures, because it is open to design choices at each level.

Looking Forward

Defining ethical goals and creating systems which distinguish right from wrong will be very difficult, but it needs to be done.

I think TalaMind will help achieve beneficial human-level AI and superintelligence faster and more safely than relying only on other methods.

However, there is much work needed to achieve human-level AI via the TalaMind approach (§7.7), e.g.:

- Create an *intelligence kernel*[9] of self-extending conceptual processes and concepts.
- Develop TalaMind's archetype / ontology level. Fully implement the linguistic level.
- Integrate the linguistic level with spatiotemporal reasoning and visualization.
- Integrate an associative level, leveraging deep neural nets, Bayesian processing.
- Develop and learn ethical concepts, encyclopedic and commonsense knowledge…
- Develop higher-level mentalities including sociality, emotional intelligence, virtues…

Acknowledgements

I thank K. Brent Venable and Vincent Conitzer for questions during the presentation of the previous paper, and David J. Kelley and Mark Waser for correspondence afterwards, motivating parts of the discussion in this paper.

Appendices

I. Introduction to the TalaMind Approach

The 'TalaMind thesis' (Jackson 2014) presents a research approach toward human-level artificial intelligence. This involves (§1.4) developing an AI system using a language of thought (called Tala) based on the unconstrained syntax of a natural language; designing this system as a collection of 'executable concepts' that can create and modify concepts, expressed in the language of thought, to behave intelligently in an environment; and using methods from cognitive linguistics such as mental spaces and conceptual blends for multiple levels of representation and computation. (Fauconnier and Turner 2002)

Proposing a design inspection alternative (§2.1) to the Turing Test, the thesis discusses 'higher-level mentalities' of human intelligence, which include natural language understanding, higher-level learning, meta-cognition and multi-level reasoning, imagination, and artificial consciousness (see Appendix II).

'Higher-level learning' (§2.1.2.5) refers collectively to forms of learning required for human-level intelligence such as learning by creating explanations and testing predictions about new domains based on analogies and metaphors with previously known domains, reasoning about ways to debug and improve behaviors and methods, learning and invention of natural languages and language games, learning or inventing new representations, and in general, self-development of new ways of thinking. The phrase 'higher-level learning' is used to distinguish these from previous research on machine learning. (Cf. Valiant 2013)

'Multi-level reasoning' refers collectively to the reasoning capabilities of human-level intelligence, including meta-reasoning, analogical reasoning, causal and purposive reasoning, abduction, induction, and deduction. (§2.1.2.6)

To provide a context for analysis of its approach the thesis discusses an architecture called TalaMind for design of AI systems (§1.5), adapted from Gärdenfors' (1995) paper on inductive inference (see Appendix III). The TalaMind architecture has three levels, called the linguistic, archetype, and associative levels. At the linguistic level, the architecture includes the Tala language, a conceptual framework for managing concepts expressed in Tala, and conceptual processes that operate on concepts in the conceptual frame-work 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 associative level would typically interface with a real-world environment and supports connectionism, Bayesian processing, etc. In general, the thesis is agnostic about research choices at the archetype and associative levels.

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

[9] A term from (Jackson 1979) corresponding to 'seed AI' (Yudkowsky 2007).

The theoretical basis for Tala is discussed in Chapter 3 of the TalaMind thesis. Section 3.3 argues it is theoretically possible to use the syntax of a natural language to represent meaning in a conceptual language and to reason directly with natural language syntax, at the linguistic level of the TalaMind architecture.

The Tala language responds to McCarthy's 1955 proposal for a formal language that corresponds to English (viz. thesis §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, conjecture, and higher-level concepts, with underspecification and semantic annotation. 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).

Chapter 3's analysis shows the TalaMind approach can address theoretical questions not easily addressed by more conventional approaches. For instance, it supports reasoning in mathematical contexts, but also supports reasoning about self-contradictory beliefs. (§3.6.6.2) 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. (§3.6.7.5) 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 thesis describes the design of a prototype demonstration system, and discusses processing in the system that illustrates the potential of the research approach to achieve human-level AI.

Of course, the thesis does not claim to actually achieve human-level AI. It only presents a theoretical direction that may eventually reach this goal, and identifies areas for future AI research to further develop the approach. These include areas previously studied by others which were outside the scope of the thesis, such as ontology, common sense knowledge, spatial reasoning and visualization, etc.

The TalaMind approach is similar though not identical to the 'deliberative general intelligence' approach proposed by (Yudkowsky 2007), as discussed in (Jackson 2014, §2.3.3.5). The architectural diagrams for human-like general intelligence given by (Goertzel, Iklé, and Wigmore 2012) may be considered as design aspects for TalaMind.

II. Artificial Consciousness

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

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

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

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

III. TalaMind's Relation to Gärdenfors (1995)

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

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

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

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

Gärdenfors' insights remain relevant, even though his discussion of the linguistic level focused on descriptions using formal languages. However, (Gärdenfors 1995) did not discuss support for the TalaMind hypotheses at the linguistic level, and did not include elements of the linguistic level discussed in the TalaMind thesis, i.e. the Tala lan-

guage, a conceptual framework for managing concepts expressed in Tala, and conceptual processes that operate on concepts in the conceptual framework to produce intelligent behaviors and new concepts. Thus (Gärdenfors 1995) did not discuss higher-level learning and other higher-level mentalities, nor aspects of minds discussed in the present paper.

References

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

Anderson, S. L. 2005. Asimov's 'Three Laws of Robotics' and machine metaethics. In (Anderson et al. 2005).

Anderson, M., Anderson, S. and Armen, C. 2005. (eds.) *Machine Ethics: Papers from the AAAI Fall Symposium*, Technical Report FS-05-06, AAAI Press.

Anderson, M. and Anderson, S. L. 2007. The status of machine ethics: A report from the AAAI Symposium. *Minds and Machines*, July 2007, 1-19.

Asimov, I. 1950. *I, Robot*. Bantam Books.

Asimov, I. 1976. *The Bicentennial Man and Other Stories*. Doubleday.

Bostrom, N. 2014. *Superintelligence – Paths, Dangers, Strategies*. Oxford University Press.

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

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

Fauconnier, G. and Turner, M. 2002. *The Way We Think – Conceptual Blending and the Mind's Hidden Complexities.* 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., Iklé, M. and Wigmore, J. 2012. The architecture of human-like general intelligence. *Foundations of Artificial General Intelligence*, 1-20.

Goff, P. 2017. *Consciousness and Fundamental Reality*. Oxford University Press.

Goff, P. 2018. Is the Universe a conscious mind? *Aeon*.

Jackson, P. C. 1979. Concept – A Context for High-Level Descriptions of Systems Which Develop Concepts, M.S. Thesis, University of California at Santa Cruz.

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. 2017. Toward human-level models of minds. *AAAI Fall Symposium Series Technical Reports*, FS-17-05, 371-375.

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

Kuipers, B. 2018. How can we trust a robot? *Communications of the ACM*, March 2018, 61, 3, 86-95.

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

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

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

Minerva, F. and Rorheim, A. 2017. What are the ethical consequences of immortality technology? *Aeon*.

Newell, A. and Simon, H. A. 1976. Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, 19, 3, 113-126.

Redd, E. and Younger, A. S. 2017. A mathematical and physical base for 'A Standard Model of the Mind'. *AAAI 2017 Fall Symposium on 'A Standard Model of the Mind'*.

Stewart, T. C. and Eliasmith, C. 2017. Continuous and parallel: challenges for a Standard Model of the Mind. *AAAI 2017 Fall Symposium on 'A Standard Model of the Mind'*.

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*. Basic Books.

Yudkowsky, E. 2007. Levels of organization in general intelligence. In *Artificial General Intelligence*, ed. B. Goertzel & C. Pennachin, 389-501.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global Catastrophic Risks*, ed. N. Bostrom & M. M. Ćircović, 308-345. Oxford University Press.