Toward Beneficial Human-Level AI… and Beyond

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

Abstract

This paper considers ethical, philosophical, and technical topics related to achieving beneficial human-level AI and superintelligence. Human-level AI need not be human-identical: The concept of self-preservation could be quite different for a human-level AI, and an AI system could be willing to sacrifice itself to save human life. Artificial consciousness need not be equivalent to human consciousness, and there need not be an ethical problem in switching off a purely symbolic artificial consciousness. The possibility of achieving superintelligence is discussed, including potential for 'conceptual gulfs' with humans, which may be bridged. Completeness conjectures are given for the 'TalaMind' approach to emulate human intelligence, and for the ability of human intelligence to understand the universe. The possibility and nature of strong vs. weak superintelligence are discussed. Two paths to superintelligence are described: The first path could be catastrophically harmful to humanity and life in general, perhaps leading to extinction events. The second path should improve our ability to achieve beneficial superintelligence. Human-level AI and superintelligence may be necessary for the survival and prosperity of humanity.

The Future of Humanity: Ethics and AI

Some potential consequences of general artificial intelligence were outlined in (Jackson 1974). Two possibilities for the "harvest of AI" were briefly discussed: A world with the machine as dictator, and a world with "well-natured machines" having enormous benefits to humanity.

Relatively recent work on 'artificial general intelligence' (Goertzel and Pennachin 2007) has included substantive research on AGI's potential consequences for humanity: Bostrom, Omohundro, Tegmark, Yudkowsky and others have discussed future scenarios in which AGI could lead to superintelligent systems with good or bad conduct toward humanity. AGI may be necessary for the survival and prosperity of humanity but if AGI is not developed very carefully it could lead to the extinction of humanity.

Ethics is the branch of philosophy that studies concepts of right and wrong (good and bad) conduct. Until recently ethics has only needed to focus on conduct by humans. Ethics and AI research now intersect regarding concepts of right and wrong conduct by intelligent machines, and right and wrong conduct in human applications of intelligent machines.

This is a challenge for AI scientists because ethical concepts of right and wrong go beyond simple questions of whether factual or theoretical knowledge is true or false, or whether problem solving behavior is successful or unsuccessful. In general, we cannot expect concepts of right and wrong conduct to be easily understood by machines. It can be a challenge for humans to distinguish these concepts sometimes.

Yet if the survival and prosperity of humanity are at stake, we are obligated to accept the challenge. Hence this paper will consider ethical, philosophical, and technical topics related to achieving beneficial human-level AI and superintelligence. The term 'beneficial' in this context does not seem to have any rigorous, agreed-upon definition. It will be used broadly to refer to consequences that are positive for humanity and biological life[1] in general.

The Possibility of Human-Level AI

A first question is whether human-level AI is even possible: The 'TalaMind thesis' (Jackson 2014) presents a research approach toward human-level artificial intelligence, which will support this paper's discussion of human-level AI's implications for the future of humanity.

The thesis endeavors to address all the major theoretical issues and objections that might be raised against its approach, or against the possibility of achieving human-level AI in principle. No insurmountable objections are identified, and arguments refuting several objections are presented. Thesis section 7.8 gives reasons in favor of the TalaMind approach over other approaches to human-level AI.

[1] Life based on DNA that has been developed by evolution. (Cf. Tegmark 2017).

The approach involves developing an AI system using a language of thought (called 'Tala') based on the unconstrained syntax of a natural language; designing the system as an 'intelligence kernel', i.e. a collection of 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 mental representation and computation.

Proposing a design inspection alternative to the Turing Test, the thesis discusses 'higher-level mentalities' of human intelligence, which include natural language understanding, higher-level learning, meta-cognition, imagination, and artificial consciousness.

'Higher-level learning' 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; Goertzel and Monroe 2017)

The thesis discusses an architecture called TalaMind for design of systems following its approach. The architecture is open, e.g. permitting predicate calculus and conceptual graphs in addition to the Tala language, and permitting deep neural nets and other methods for machine learning.

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.

Human-Level AI ≠ Human-Identical AI

The TalaMind thesis gives reasons why the Turing Test does not serve as a good definition of the goal we are trying to achieve, human-level AI. In particular, the Turing Test conflates human-level intelligence with human-identical intelligence, i.e. intelligence indistinguishable from humans. This is important because in seeking to achieve human-level AI we need not seek to replicate human thinking. Human-level AI can be 'human-like' without being human-identical. (Jackson 2014, §2.1.1)

In particular for beneficial AI, the concept of self-preservation could be quite different for a human-level AI than it is for a human. A human-level AI could periodically backup its memory, and if it were physically destroyed, it could be reconstructed and its memory restored to the backup point. So even if it had a goal for self-preservation, a human-level AI might not give that goal the same importance a human being does. It might be more concerned about protection of the technical infrastructure for the backup system, which might include the cloud, and by extension, civilization in general.

A human-level AI could understand that humans cannot backup and restore their minds, and regenerate their bodies if they die, at least with present technologies. It could understand that self-preservation is more important for humans, than for AI systems. The AI system could be willing to sacrifice itself to save human life, especially knowing that as an artificial system it could be restored.

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.

Symbolic Artificial Consciousness ≠ Human Consciousness

The axioms of artificial consciousness can be implemented with symbolic representations and symbolic processing, as illustrated in the TalaMind prototype. The human first-person subjective experience of consciousness is much richer and more complex. Achieving human-level AI may not require achieving human-identical consciousness in an AI system.

This is important to note because some authors seem to assume artificial consciousness will be equivalent to human consciousness, and assume a system with artificial consciousness should automatically have the same moral status and legal protections as a human being, so that switching off the system could be immoral or illegal. Some even suggest that if a system simulates consciousness within itself, and then terminates the simulation, the system may have performed a 'mind crime'. (Bostrom 2014)

Such suggestions are at best philosophical, and at worst Orwellian, if a system with symbolic artificial consciousness does not have any subjective experiences approaching human consciousness. Switching off such a system is not worse than switching off any computer that performs symbolic processing. Whether it is right or wrong to stop such a system depends on whether its symbolic processing would 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.

Further, to support reasoning about potential future events, and counterfactual reasoning about past and present events, a system may need to simulate what other intelligent systems and people may think or do, and then terminate its simulations. The TalaMind thesis (Jackson 2014) uses the term 'nested conceptual simulation' to refer to an agent's conceptual processing of hypothetical scenarios, with possible branching of scenarios based on alternative events, such as choices of simulated agents. This amounts to a Theory of Mind capability within a TalaMind architecture, 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 will be necessary for human-level AI, and should not be considered mind crimes.

For the same reason, relying on robots with such limited, symbolic artificial consciousness is not a form of 'slavery'. It is just symbolic processing.

The Possibility of Superintelligence

Since one of the abilities of human intelligence is the ability to design and improve machines, it's natural to suppose human-level AI could be applied to improve itself, and to think this might lead to "runaway" increases in machine intelligence beyond the human level. This possibility was first suggested[2] by Good (1965), and later considered by Vinge (1993), Moravec (1998), Kurzweil (2005), and others. Bostrom (2014) and Tegmark (2017) give current discussions.

To evaluate whether superintelligence can be achieved, let's consider what it could mean to "improve" human-level artificial intelligence, and whether and how human-level AI could improve itself to achieve superintelligence.

Here's a list of ways human-level AI could be improved relative to human intelligence:

Sensory capabilities – An AI system could perceive light (and sound) at different wavelengths, and phenomena at different scales (smaller or larger) than humans can directly observe.

Active capabilities – An AI system could perform actions at different physical scales than humans can directly perform.

Speed of thought – A computer can perform logical operations at speeds orders of magnitude faster than a neuron can fire. This may translate to corresponding speedups in thought.

Information access – An AI system could in principle access all the information in Wikipedia, or even the entire Web. A human-level AI could understand much of this information.

Extent and duration of memory – An AI system could in principle remember everything it has ever observed. Only a few humans claim this ability.

Duration of thought – A human-level AI could continue thinking about a particular topic for years, decades, centuries, millennia, … .

Community of thought – A collection of human-level AI's could share thoughts (conceptual structures) more directly, more rapidly, and less ambiguously than a collection of humans. If human-level AI can be copied and processed inexpensively, then much larger groups of

[2] Two earlier related suggestions are noteworthy: Turing (1950) asked "Can a machine be made to be super-critical?" i.e. to generate ideas in a manner analogous to super-criticality of nuclear reactions. Ulam (1958) recalled a conversation with von Neumann "on the ever accelerating progress of technology…which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue."

human-level AI's could be assembled to collaborate on a topic than would be possible with humans.[3]

Nature of thought – A human-level AI (or community of HLAI's) can develop new concepts and new conceptual processes. Such concepts and processes may be developed more rapidly than humans develop or understand them, creating *'conceptual gulfs'* in understanding between AI systems and humans.

Recursive self-improvement – This term does not seem to have any rigorous, agreed-upon definition though it is frequently used to describe how superintelligence could be achieved. Essentially it could be the recursive compounding of all the above improvement methods, and any other specific methods which may be identified.

These characteristics might all be described as 'more and faster' human-level AI, and may be called 'weak' superintelligence (cf. Vinge 1993). If human-level AI is achieved then it will be possible to create weak superintelligence.

Completeness of the TalaMind Approach

In effect, the TalaMind thesis (Jackson 2014) conjectures that the extensible 'nature of thought' for a TalaMind architecture is complete for supporting human-level AI, since it includes concepts represented in natural language as well as mathematically and logically in formal languages, supported by conceptual levels for cognitive concept structures and associative processing, with future extensions for spatial reasoning and visualization, etc. The Tala language is also a simple universal programming language for representing executable concepts and conceptual processes. In principle, TalaMind architectures could be extended to include human-level subjective consciousness, though that is a topic for a separate, future paper. This paper focuses only on the potential for AI systems with symbolic artificial consciousness, as discussed above.

The nature of thought for human intelligence is very powerful and extensible: It has enabled Homo sapiens to transition from "an unexceptional savannah-dwelling primate to become the dominant force on the planet" (Harari 2015). This transition has leveraged the expressive power and extensibility of human natural languages, which have enabled Sapiens to represent and communicate thoughts in domains of objective knowledge about the world such as physics and biology, and intersubjective knowledge about concepts invented by humans, such as money, corporations, ethical concepts, laws, nations, etc.

Although humans have cognitive biases and individual limitations, it may not be hubris to conjecture human intelligence is completely general. Consider that scientists and mathematicians have extended human concepts into new domains not directly observed, conceptualizing multiple dimensions, universal computation, general relativity, quantum theory, etc. If human intelligence is completely general then humans may eventually understand all the phenomena in the universe, by combining abilities to invent and represent hypothetical concepts about the universe with abilities to scientifically test hypotheses – if all the phenomena in the universe can be explained by practically testable theories. That's a big "if" of course.

If the TalaMind approach can achieve human-level AI, then a completeness conjecture for human intelligence extends to the TalaMind approach, and to superintelligent systems using TalaMind architectures.

Getting Over Conceptual Gulfs

Conceptual gulfs happen normally between human minds: For example, scientists have developed concepts that are not understood by the average person, or even by scientists in other fields. The worldwide scientific community may be considered superintelligent relative to any individual human. People accept this form of superintelligence because they believe scientific ideas can be understood and validated between scientists, and they believe scientific knowledge in general is beneficial to humanity.

Likewise, conceptual gulfs between weak superintelligence and humans could be bridged and new concepts could be explained to humans. This will be facilitated if AI systems follow the TalaMind approach, using a language of thought based on a natural language. Conceivably, conceptual gulfs between weak superintelligence and humans may have short duration in some domains, though there may always be conceptual gulfs to bridge.

Is 'Strong' Superintelligence Possible?

Could a strong superintelligence exist, qualitatively superior to weak superintelligence, i.e. superior to 'more and faster' human-level AI?

The answer seems to depend on other limits and characteristics of human intelligence that are not yet known by scientists. For instance, it appears not yet known for certain whether human intelligence requires super-Turing computation or quantum computation. Even if Penrose and Hameroff's "Orch-OR" hypothesis is disproved, the possibility may remain that other forms of nanoscale quantum computation occur within the brain. Neuroscientists may

[3] The TalaMind hypotheses do not require a society of mind architecture, but it is natural to implement a society of mind at the linguistic level of a TalaMind architecture. A society of mind architecture could also support a community of thought for human-level AI's. (Cf. Jackson 2014, §2.3.3.2.1)

consider this unlikely, but so far as I know it has not been completely ruled out. The same situation may hold for super-Turing computation.

If these forms of computation are required by the brain to support human intelligence, then human-level AI would need to include them to match the abilities of human intelligence. If human intelligence is also completely general, then no stronger form of intelligence would exist other than 'more and faster' human-level intelligence, i.e. weak superintelligence.

On the other hand, if these forms of computation are not used by the brain then extending human-level AI to use them could yield a 'strong' superintelligence, able to solve some problems that would be intractable for 'more and faster' human-level intelligence. Likewise, if human intelligence is not completely general then making human-level AI completely general could yield a strong superintelligence surpassing 'more and faster' human-level intelligence.

In either case, conceptual gulfs between humans and strong superintelligence could be bridged at least to the extent of using natural language to give descriptions of concepts developed by strong superintelligence.

Two Paths to Superintelligence

There are at least two somewhat different paths toward superintelligence. One path would focus on recursive self-improvement of general AI systems (AGI) having unchangeable 'final goals' which may be relatively simple and arbitrary. Bostrom (2014) discussed several ways this path could achieve superintelligence that would be catastrophically harmful to humanity and life in general, perhaps leading to extinction events.

Yudkowsky (2008) noted the design space for AGI is much larger than human intelligence, writing "The term 'Artificial Intelligence' refers to a vastly greater space of possibilities than does the term 'Homo sapiens.'" He strongly urged readers not to assume a fully general optimization process for AGI will be beneficial to humanity, yet advised not writing off the challenge of beneficial AI.

A second path toward superintelligence, consistent with the TalaMind approach, focuses on limiting the research design space to AI systems which have generality and which also have higher-level mentalities that are characteristic of human intelligence. This design space would be further limited to systems for which the only unchangeable goals are ethical goals beneficial to humanity and to biological life in general. This narrowing of the design space should improve our ability to achieve beneficial human-level AI and beneficial superintelligence via recursive self-improvement.

Human-Level Intelligence & Goals

In discussing the first path to superintelligence, Bostrom[4] (2014) relied on an 'orthogonality thesis' that "intelligence and final goals are independent variables: any level of intelligence could be combined with any final goal." He wrote:

> "There is nothing paradoxical about an AI whose sole final goal is to count the grains of sand on Boracay, or to calculate the decimal expansion of pi, or to maximize the total number of paperclips that will exist in its future light cone. In fact, it would be easier to create an AI with simple goals like these than to build one that had a human-like set of values and dispositions."

In taking the second path to superintelligence, these would not be allowed as unchangeable final goals. A TalaMind system would realize it is pointless to count the grains of sand on Boracay, impossible to fully calculate the infinite decimal expansion of pi, and harmful to humanity to maximize the number of paperclips in its future light cone. So it would reject or abandon these simple goals.

Bostrom (2014) also relied on an 'instrumental convergence thesis' that "superintelligent agents having any of a wide range of final goals will nevertheless pursue similar intermediary goals because they have common instrumental reasons to do so." In particular, he cited two instrumental goals which could cause superintelligent systems to be very harmful to humanity, perhaps leading to an extinction event. The first is a goal of self-preservation. The second is a goal of maximizing available resources. I've described above how a human-level AI could have a different concept of self-preservation, facilitating self-sacrifice to save human life. This could apply also to a superintelligence.

In scenarios (Bostrom 2014) discussed, the goal of maximizing resources causes a superintelligent system to accumulate as much money and power as possible, leading to very harmful consequences for humanity. This is another case where the ability to think ethically about goals, and change or abandon them is important. A human-level AI should understand there are appropriate and inappropriate relationships between goals and possible means to achieve goals. It should understand that achieving an important goal does not justify acquiring as much money and power as possible – rather, it should have an ethical meta-goal to achieve its goals with as little resources and money as possible, and without acquiring power over human lives or human decisions.

Taking the second path won't be easier than the first path just because the design space is smaller. Framing ethical goals and creating human-level AI systems which distinguish right from wrong conduct will be very difficult,

[4] Bostrom (2014) consolidated research on the first path by himself and others, including Omohundro and Yudkowsky.

but it needs to be done. TalaMind's use of a natural language mentalese will facilitate representing ethical concepts and goals.

To achieve beneficial AI it's also important to develop the TalaMind approach because a system that reasons in a conceptual language based on English (or some other common natural language) will be more open to human inspection than a black box or a system with an internal language that's difficult for people to understand.

Looking Forward

Human-level AI and superintelligence could help develop scientific knowledge more rapidly than possible through human thought alone, and help advance medicine, agriculture, energy systems, environmental sciences, and other areas of knowledge directly benefitting human prosperity and survival.

Human-level AI may be necessary for the long-term survival and prosperity of humanity: People are not biologically suited for lengthy space travel with present technologies. To avoid depleting the Earth's resources and to avoid the fate of the dinosaurs (whether from asteroids or super-volcanoes) our species will need economical, self-sustaining settlements off the Earth. Human-level AI may be necessary for mankind to spread throughout the solar system, and later the stars.

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." We have travelled far over six decades, and can now see a path toward beneficial superintelligence.

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.

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

Bello, P. and Bringsjord, S. 2013. On how to build a moral machine. *Topoi*, 32, 2, 1-25.

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.

Doyle, J. 1983. A Society of Mind – multiple perspectives, reasoned assumptions, and virtual copies. *Proceedings 1983 International Joint Conference on Artificial Intelligence*, 309-314.

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

Goertzel, B. and Pennachin, C. eds. 2007. *Artificial General Intelligence*. Springer.

Goertzel, B., Iklé, M. and Wigmore, J. 2012. The architecture of human-like general intelligence. *Foundations of Artificial General Intelligence*, 1-20.

Goertzel, B. and Monroe, E. 2017. Toward a general model of human-like general intelligence. *AAAI Fall Symposium Series Technical Reports*, FSS-17-05, 344-347.

Good, I. J. 1965. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, vol. 6, 1965.

Hameroff, S. and Penrose, R. 2014. Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews*, 11, 1, 39 – 78. Elsevier.

Harari, Y. N. 2015. *Sapiens: A Brief History of Humankind*. HarperCollins Publishers.

Jackson, P. C. 1974. *Introduction to Artificial Intelligence*. New York: Mason-Charter Publishers.

Jackson, P. C. 1985. *Introduction to Artificial Intelligence*, Second Edition. New York: Dover Publications.

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*, FSS-17-05, 371-375.

Kurzweil, R. 2005. *The Singularity Is Near: When Humans Transcend Biology*. Viking.

Moravec, H. P. 1998. *Robot: Mere Machine to Transcendent Mind*. Oxford University Press.

Omohundro, S. M. 2008. The basic AI drives. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, ed. P. Wang, B. Goertzel & S. Franklin, 483-492.

Scheutz, M. 2017. The case for explicit ethical agents. *AI Magazine*, 38, 4, 57-64.

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

Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, 59, 433 - 460.

Ulam, S. 1958. Tribute to John von Neumann, *Bulletin of the American Mathematical Society*, 64, 3, 1 - 49.

Valiant, L. G. 2013. *Probably Approximately Correct – Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books.

Vinge, V. 1993. The coming technological singularity: how to survive in the post-human era. *Whole Earth Review*, Winter 1993.

Walsh, T. 2017. The singularity may never be near. *AI Magazine*, 38, 3, 58 - 62.

Wilks, Y. 2017. Will there be superintelligence and would it hate us? *AI Magazine*, 38, 4, 65-70.

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.