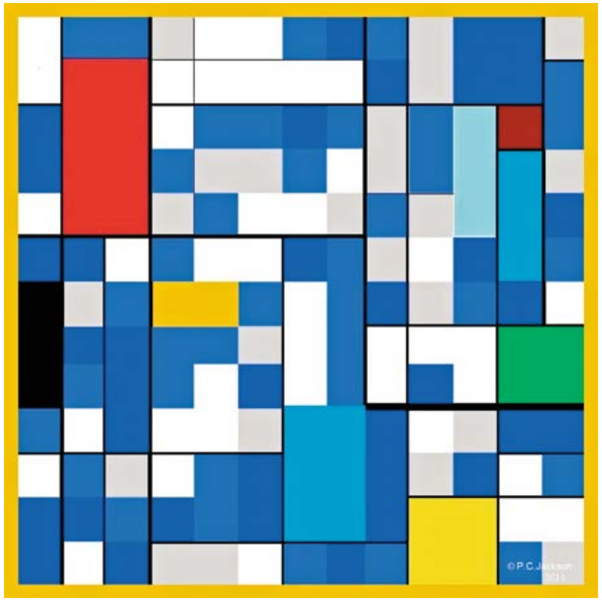
Toward Beneficial Human-Level AI… and Beyond

Philip C. Jackson, Jr.
TalaMind LLC
www.talamind.com



AAAI Spring Symposium
March 26-28, 2018
Stanford University

Topics

- The Future of Humanity and AI
- The Possibility of Human-Level AI: TalaMind
- Human-Level AI ≠ Human-Identical AI
- TalaMind Architecture / Research Direction
- Natural Language and Ethical Concepts
- Artificial Consciousness
- Symbolic Artificial Consciousness ≠ Human Consciousness
- A Counter-Argument
- Acting As If Robots are Fully Conscious
- Human-Level AI and Self-Preservation
- Avoiding Artificial Slavery
- Theory of Mind and Simulations of Minds
- A Mind is a Universe unto Itself
- …and Beyond: The Possibility of Superintelligence
- Nature of Thought and Conceptual Gulfs
- Is 'Strong' Superintelligence Possible?
- Two Paths to Superintelligence
- Superintelligence & Final Goals
- Superintelligence & Instrumental Goals
- Taking the Second Path, Toward Beneficial AI and Beyond

I will go quickly through these topics and have about 10 minutes for questions after the slides. Will be glad to discuss with people afterwards.

Some of these topics go beyond the present paper, and are discussed in a forthcoming 'postscript' paper. These slides are posted at www.talamind.com.

The Future of Humanity and AI

Two potential scenarios for '*the harvest of AI*' were outlined in (Jackson 1974):

- A world with the machine as dictator.
- A world with "well-natured machines" having enormous benefits to humanity.

Research on *'artificial general intelligence'* has included substantive arguments* that if AGI is not developed carefully it could be catastrophically harmful to humanity.

Yet beneficial AI may be needed for the survival and prosperity of humanity:

- AI may enable universal basic income, global economic growth, eliminating poverty.
- We may need human-level AI to help settle the Solar System and explore the stars.

So, we are obligated to consider the problem:

How to insure human-level AI and superintelligence will be beneficial to humanity?

* Bostrom, Omohundro, Yudkowsky, Tegmark, and others.

The Possibility of Human-Level AI: TalaMind

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.
  - It supports this paper’s discussion of human-level AI’s implications for humanity’s future.
  - It’s important for achieving beneficial human-level AI, for reasons I’ll discuss.
- The thesis discusses theoretical issues and objections against its approach, or against achieving human-level AI in principle.
- No insurmountable objections are identified. (§4.3)*

*The notation §4.3 refers to Chapter 4, section 3 of the TalaMind thesis (Jackson 2014).

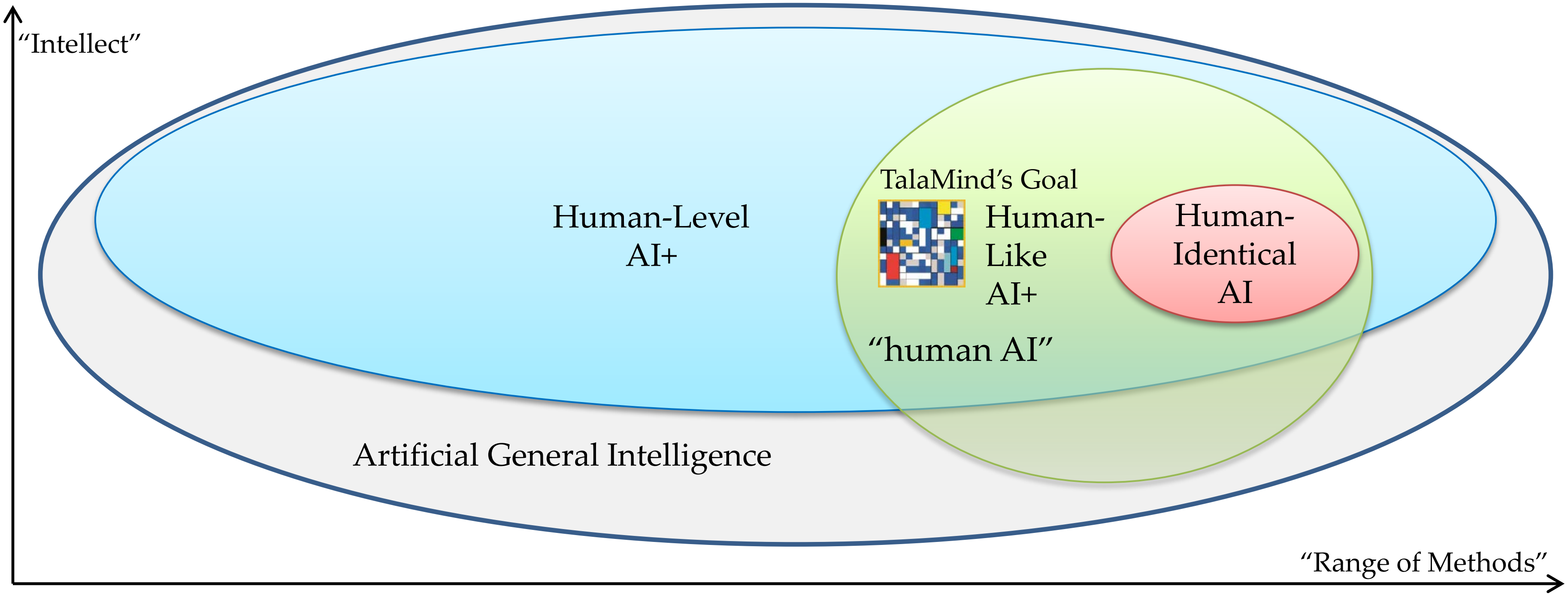
Human-Level AI ≠ Human-Identical AI

Human-level AI can be 'human-like' without being human-identical.

AI can use thought processes similar to humans, and understandable by humans.
The TalaMind approach strives for this.

TalaMind defines human-level AI by a *design inspection approach* (Jackson 2014, §2.1.1) :

- Achieving *higher-level mentalities* which people agree indicate human-level intelligence.



TalaMind Architecture / Research Direction

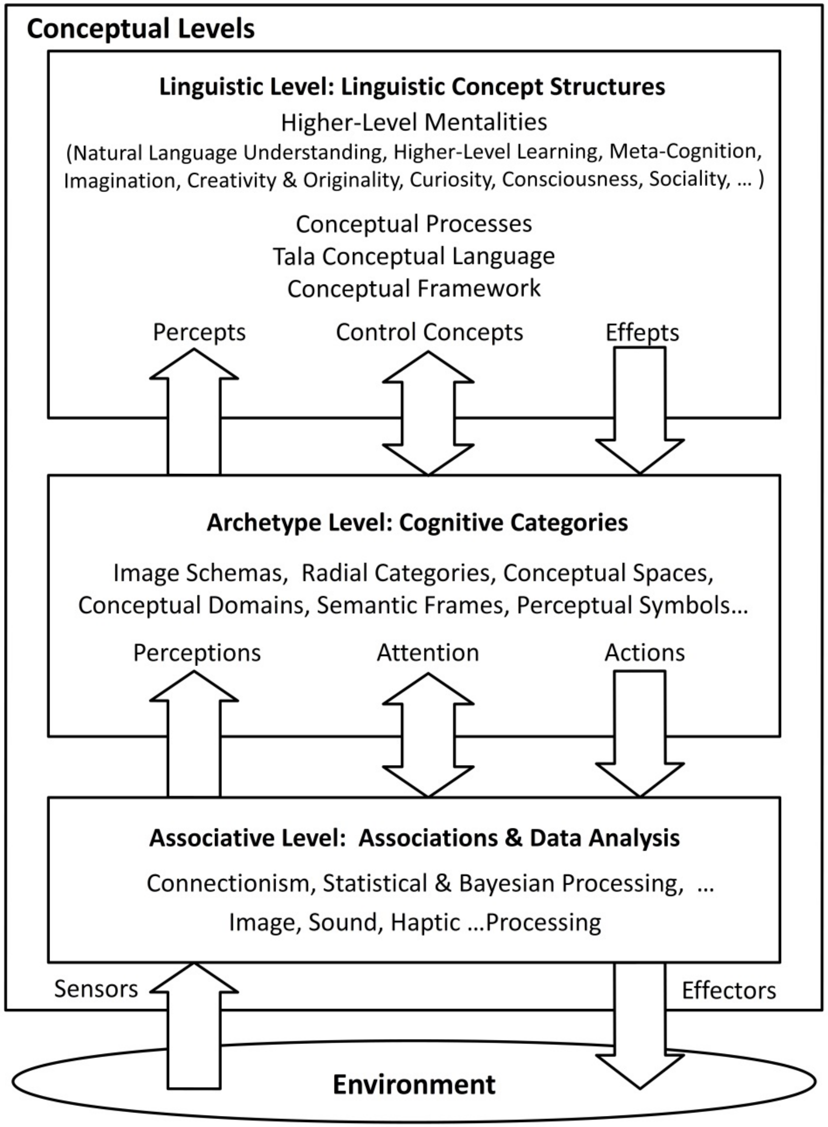
At the linguistic level, TalaMind:

- has a language of thought (called 'Tala') with the unconstrained syntax of English.
- supports reasoning directly with the syntax of English.
- argues this is theoretically valid, and advantageous for human-level AI.

Tala (the natural language mentalese):

- will facilitate representing ethical concepts.
- will be understandable to humans and open to human inspection

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


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

TalaMind is open to use of additional languages, e.g. predicate calculus and conceptual graphs.

TalaMind is open to use of deep neural nets for each conceptual level.

Natural Language and Ethical Concepts

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

TalaMind will support a recommendation that ethical principles for robots should be expressed in natural language:

> "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 would do more: It would 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, …

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.

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

Artificial Consciousness

The TalaMind thesis accepts the objection that a system which is not aware of what it is doing and does not have some awareness of itself cannot have human-level intelligence.

The TalaMind thesis adapts the “axioms of being conscious” proposed by Aleksander and Morton (2007). 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.

Symbolic Artificial Consciousness ≠ Human Consciousness

The axioms of artificial consciousness can be implemented with symbolic representations and symbolic processing.

The human first-person subjective experience of consciousness is richer and more complex, though we don't know precisely how to explain it.

Halting a symbolic consciousness may not be worse than halting 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 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, there is a counter-argument.

A Counter-Argument

Arguably, Newell & Simon's (1976) Physical Symbol System Hypothesis implies a purely symbolic artificial consciousness could be equivalent to human consciousness.

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

This argument may be valid: We don't know how to fully explain human consciousness and there may be some form of symbolic processing that's equivalent to human consciousness.

For discussion, I'll call this "artificial subjective consciousness".

It has not been proved that computers cannot achieve all the capabilities of the human brain including human-level subjective consciousness. (§4.3)

Still, artificial subjective consciousness would be more complex than the axioms for symbolic artificial consciousness (Aleksander and Morton 2007).

Symbolic Artificial Consciousness << Artificial Subjective Consciousness

Newell & Simon's definition of a physical symbol system evidently includes neural nets, and any methods that can be implemented on a computer.

The TalaMind approach does not appear to be in conflict with eventually achieving artificial subjective consciousness, if that is possible. (§4.2.7)

Acting As If Robots Are Fully Conscious

Apart from whether AI systems actually achieve human-level consciousness, there are ethical arguments 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. (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 needed for human-level consciousness.

Yet 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.

Anderson, S. L. 2005. Asimov's 'Three Laws of Robotics' and machine metaethics. In Anderson, M., Anderson, S. and Armen, C. 2005. (eds.) *Machine Ethics: Papers from the AAAI Fall Symposium*, Technical Report FS-05-06, AAAI Press.

Human-Level AI and Self-Preservation

Because human-level AI ≠ human-identical AI,

the concept of self-preservation can be different for a human-level AI than it is for a human.

An AI system can backup its memory, and if it is physically destroyed, it can be reconstructed and its memory restored to the backup point.

Even if it has a goal for self-preservation, an AI system might not give that goal the same importance a human being does.

It could understand humans cannot backup and restore their minds, and regenerate their bodies if they die, at least with present technologies.

It could understand self-preservation is more important for humans than for AI systems.

So an AI system could be willing to sacrifice itself to save human life.

Avoiding Artificial Slavery

Even if artificial subjective consciousness is achieved, relying on such systems is not inherently equivalent to slavery:

Artificially conscious systems could have goals to be beneficial to humanity, yet not be slaves.

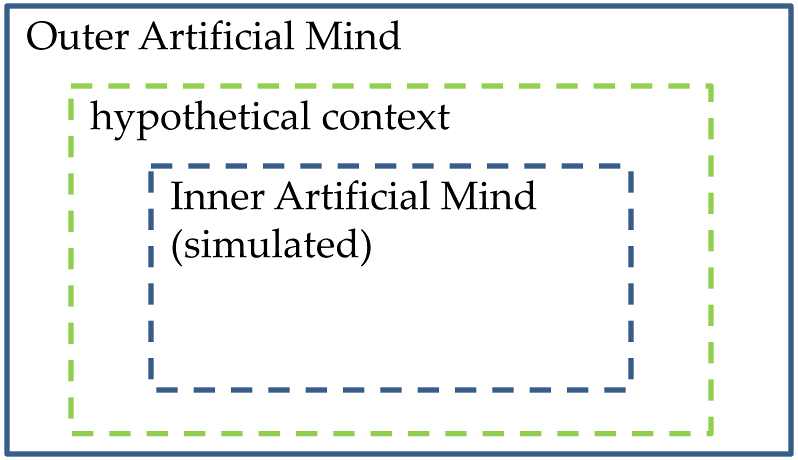
Human-level robots could have autonomy and independence in choosing how to be beneficial, whom to work with or work for, etc.

They may consider themselves as extensions of humanity. We may think of them that way, also.

Theory of Mind and Simulations of Minds

To reason about past, present and potential future events, a system may need to simulate what other systems and people may think or do.

So an artificial mind may need to simulate other minds and halt its simulations.



This supports a 'Theory of Mind' ability for an AI system.

Arguably, such simulations will be necessary for human-level AI.

However, some authors have suggested that if an artificial mind simulates another mind and then halts the simulation, the system may have committed a 'mind crime'. (Bostrom 2014)

A Mind is a Universe unto Itself

I propose taking an ethical and philosophical stance that a mind may be considered as a universe unto itself.

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 may be similar to dreaming. The mind can stop a dream, halting simulation of imaginary actors.

In this ethical stance, artificial minds have a degree of freedom of thought and control of thought within their individual scopes.

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

This philosophical stance does not contend our physical Universe is itself a mind or is governed by a mind. Goff discusses how this may be implied by what is known about the laws of physics.

Goff, P. 2017. *Consciousness and Fundamental Reality*. Oxford University Press.
Goff, P. 2018. Is the Universe a conscious mind? Aeon.

A Mind is a Universe unto Itself, II

Arguably, to avoid an ethical problem if an artificial mind simulates and halts minds with artificial subjective consciousness,

> the outer mind might only create internal simulations of itself in situations it envisions for other minds.

Typically this may be the most any mind can do anyway in trying to understand other minds.

Such simulations may help an artificial mind support empathy for other minds – though empathy also requires understanding emotions and ethical concepts (e.g. fairness).

An ethical problem may also be avoided if an outer mind only reasons about what other minds might think and feel, without simulating artificial subjective consciousness of other minds.

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

This ethical stance would still be problematical if an artificial mind were to internally simulate a living human mind that was uploaded to a computer.

Uploading Human Consciousness

Future technologies may be able to scan the neurons in a human brain and replicate neural processing within a computer (Markram 2006).

This may give us a much better understanding of what human consciousness is.

If these technologies can be developed, it could give human minds near-immortality and freedom from paralyzed or dying bodies.

This would raise a host of ethical questions about uploading minds and human immortality – too many to discuss or even list them here.

Yet I will give some brief thoughts:

Arguably, uploaded human minds should be given similar protections to biological human minds, but not greater protections.

Biological human minds would be more evanescent and may need greater protections.

AI Simulations of Uploaded Human Minds

To prevent situations where an artificial mind might upload a human mind, simulate the human mind within itself, and then halt the simulation:

We could hold that every human mind has a unique copyright to itself and to its human brain.

We could give AI systems ethical rules regarding 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 in its original living brain or as an uploaded mind that is autonomous and not simulated within another system.

… and Beyond: The Possibility of Superintelligence

It's natural to think AI could improve itself and that this might lead to "runaway" increases in intelligence beyond the human level.

- This was discussed by Good (1965), and later considered by Vinge (1993), Moravec (1998), Kurzweil (2005), and others. Bostrom (2014) and Tegmark (2017) give current discussions. The possibility was suggested by Turing and von Neumann separately in the 1950's.

There are several ways AI could be improved relative to human intelligence:

> Sensory capabilities, Active capabilities, Speed of thought, Information access , Extent and duration of memory, Duration of thought , Community of thought, Nature of thought,

Recursive self-improvement: the recursive compounding of improvement methods.

Systems with these improvements might be described as 'more and faster' human-level AI, and called 'weak' superintelligence (cf. Vinge 1993).

If human-level AI is achieved then it will be possible to create weak superintelligence.

> The TalaMind approach will help achieve superintelligence: Its natural language mentalese could be used for 'communities of thought' by human-level AI's.

Nature of Thought and Conceptual Gulfs

Nature of thought – A human-level AI can develop new concepts and new conceptual processes, e.g. new ways of solving problems.

TalaMind’s natural language mentalese will support developing new concepts and new conceptual processes, arguably better than formal logical languages.

New concepts and processes may be developed more rapidly than humans develop or understand them, creating ‘conceptual gulfs’ between AI systems and humans.

Conceptual gulfs happen normally between human minds:

For example, scientists have developed concepts not understood by the average person, or by scientists in other fields.

The worldwide scientific community is superintelligent relative to any individual human.

Conceptual gulfs between weak superintelligence and humans could be bridged and new concepts could be explained to humans.

This will be facilitated by TalaMind’s language of thought based on English.

Is ‘Strong’ Superintelligence Possible?

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

My paper discusses how this question could be answered yes or no depending on limits and characteristics of human intelligence that are not yet known by scientists.

If strong superintelligence can exist then conceptual gulfs with humans could still be bridged, at least partially, by using natural language to describe 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.*

A second path, consistent with the TalaMind approach, focuses on limiting the research design space to human-like AI systems:

- which are understandable to humans and open to human inspection, and

- for which the paramount 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.

*Bostrom cited research by Omohundro, Yudkowsky, and others.

Superintelligence & Final Goals

In discussing the first path to superintelligence, Bostrom (2014) described some simple, at first-glance harmless goals that could lead to disasters.

In taking the second path to superintelligence, these would not be allowed as unchangeable final goals.

An ethical AI+ system could realize many simple goals are pointless, impossible, or harmful if taken to extremes:

- pointless to count the grains of sand on a beach,
- impossible to calculate all the digits of pi,
- harmful to maximize paperclips in its future light cone.

So it would reject or abandon these simple goals.

In addition to ethics, rejecting such goals involves knowledge about physics, mathematics, economics, etc.

Superintelligence & Instrumental Goals

Bostrom (2014) also discussed ‘instrumental goals’ which could be harmful to humanity:

- Self-preservation.
  - I’ve described above how an AI system could have a different concept of self-preservation, facilitating self-sacrifice to save human life.
- Maximizing available resources.
  - In taking the second path, an AI system should have an ethical meta-goal to achieve its goals with minimal resources, i.e. as little money and power as possible.

Taking the Second Path, Toward Beneficial Human-Level AI… and Beyond

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

- Create an *intelligence kernel** 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…

*A term from (Jackson 1979) corresponding to 'seed AI' (Yudkowsky 2007).