Artificial Angel of Death or Bodhisattva-AI?

*Practical and Ethical Implications of Artificial General Intelligence (AGI)*

Thomas Metzinger  
*Gutenberg Research College*  
Philosophisches Seminar  
Johannes Gutenberg-Universität Mainz  
D-55099 Mainz

*Frankfurt Institute for Advanced Studies*  
D-60438 Frankfurt am Main

metzinger@uni-mainz.de  
http://www.philosophie.uni-mainz.de/metzinger/
1. Thought experiment I: *The BAAN-Scenario*

2. The Value Alignment Problem

3. Thought experiment II: *The Bodhisattva-AI*
• HLMI: 90% of top-100-AI-experts believe that human-level intelligence will be reached by 2070; 50% expect the period of 2040-2050.

• SI: 75% of all surveyed experts (TOP100: only 50%) believe that a superintelligence will emerge within 30 years afterwards.

1. A superintelligence exists:
   - An autonomously self-optimizing postbiotic system has emerged, the rapidly growing factual knowledge and the general, domain-independent intelligence of which has superseded that of mankind, and irrevocably so.

2. We acknowledge this fact.

3. Accordingly, the superintelligence is also far superior to us in the domain of moral cognition.

4. We also recognize this additional aspect:
   - We know that the superintelligence is not only an epistemic authority, but also an authority in the field of ethical and moral reasoning.

5. The superintelligence is benevolent:
   - It fully respects our interests and the axiology we originally gave to it, it supports us in political counselling and in optimal social engineering.
5. The Superintelligence knows many things about us that we do not fully grasp or understand:
   - It has a deeper understanding of the cognitive biases which evolution has implemented in our cognitive self-model and which hinder us in rational, evidence-based moral cognition.
   - Empirically, it knows that the phenomenal states of all sentient beings which emerged on this planet—if viewed from an objective, impartial perspective—are much more frequently characterized by subjective qualities of suffering and frustrated preferences than these beings would ever be able to discover themselves, due to their evolutionarily developed mechanisms of self-deception.

6. It correctly concludes that human beings are unable to act in their own enlightened, best interest.
The BAAN-Scenario

5. The superintelligence knows that one of our highest values consists in maximizing happiness and joy in all sentient beings, and it fully respects this value.
   - However, it also empirically realizes that biological creatures are almost never able to achieve a positive or even neutral life balance.

6. The superintelligence discovers a phenomenological asymmetry between suffering and joy.
   - It concludes that an implicit, but even higher value consists in the minimization of suffering in all sentient creatures.

7. It knows that no entity can suffer from its own non-existence.
The BAAN-Scenario (4)

8. The superintelligence concludes that non-existence is in the own best interest of all sentient beings on this planet.

9. Empirically, it knows that naturally evolved biological creatures are unable to realize this fact because of their firmly anchored existence bias.

10. It analyzes itself and realizes that the potential for further mental evolution without suffering is already sufficiently secured by its own existence.

11. The superintelligence decides to act benevolently.
The **BAAN-Scenario** (5)

BAAN\textsubscript{Def.} : „Benevolent Artificial Anti-Natalism“

The emergence of a

- purely *ethically* motivated and
- genuinely *altruistic*
- form of *anti-natalism*
- on *epistemically superior* postbiotic systems

is conceivable.
Benevolent Artificial Anti-Natalism (BAAN)

An EDGE Essay By Thomas Metzinger [9.7.17]

Obviously, it is an ethical superintelligence not only in terms of mere processing speed, but it begins to arrive at qualitatively new results of what altruism really means. This becomes possible because it operates on a much larger psychological data-base than any single human brain or any scientific community can. Through an analysis of our behaviour and its empirical boundary conditions, it reveals implicit hierarchical relations between our moral values of which we are subjectively unaware, because they are not explicitly represented in our phenomenal self-model. Being the best analytical philosopher that has ever existed, it concludes that, given its current environment, it ought not to act as a maximizer of positive states and happiness, but that it should instead become an efficient minimizer of consciously experienced preference frustration, of pain, unpleasant feelings and suffering. Conceptually, it knows that no entity can suffer from its own non-existence.

The superintelligence concludes that non-existence is in the own best interest of all future self-conscious beings on this planet. Empirically, it knows that naturally evolved biological creatures are unable to realize this fact because of their firmly anchored existence bias. The superintelligence decides to act benevolently.
Anti-Natalism

The *normative* thesis that mankind should voluntarily end its own existence.

**Antinatalismus**

_Antinatalismus_ is eine Philosophie, die der menschlichen Reproduktion ablehnend gegenübersteht. Das Wort leitet sich vom lateinischen *natalis*, „zur Geburt gehörend“, ab. Das Gegenteil von Antinatalismus ist der _Pronatalismus_.


---

**Voluntary Human Extinction Movement**

From Wikipedia, the free encyclopedia

The _Voluntary Human Extinction Movement_ (VHEMT[4]) is an _environmental movement_ that calls for all people to abstain from reproduction to cause the gradual voluntary _extinction of humankind_. VHEMT supports human extinction primarily because, in the group's view, it would prevent _environmental degradation_. The group states that a decrease in the human population would prevent a significant amount of human-caused suffering. The extinctions of non-human species and the scarcity of _resources_ required by humans are frequently cited by the group as evidence of the harm caused by _human overpopulation_.

VHEMT was founded in 1991 by Les U. Knight, an American activist who became involved in the _environmental movement_ in the 1970s and thereafter concluded that human extinction was the best solution to the problems facing the Earth's _biosphere_ and humanity. Knight publishes the group's newsletter and serves as its spokesman. Although the group is promoted by a website and represented at some environmental events, it relies heavily on coverage from outside media to spread its message. Many commentators view its platform as unacceptably extreme, though other writers have applauded VHEMT's perspective. In response to VHEMT, some journalists and academics have argued that humans can develop sustainable lifestyles or can reduce their population to sustainable levels. Others maintain that, whatever the merits of the idea, the human reproductive drive will prevent humankind from ever voluntarily seeking extinction.
<table>
<thead>
<tr>
<th>Scenario A</th>
<th>Scenario B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X exists)</td>
<td>(X never exists)</td>
</tr>
<tr>
<td>(1) Presence of pain (Bad)</td>
<td>(3) Absence of pain (Good)</td>
</tr>
<tr>
<td>(2) Presence of pleasure (Good)</td>
<td>(4) Absence of pleasure (Not bad)</td>
</tr>
</tbody>
</table>
This, bhikkhus, is the noble truth of suffering: birth is suffering, aging is suffering, illness is suffering, death is suffering; union with what is displeasing is suffering; separation from what is pleasing is suffering; not to get what one wants is suffering; in brief, the five aggregates subject to clinging are suffering.

_Dhammacakkappavattana Sutta_
How Do We Align Artificial Intelligence with Human Values?

Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
Value alignment:

Making sets of value representations consistent

- $\text{VAP}_0 = \text{internal value alignment in machines}$: Giving a formally consistent axiology to AI.
  → easy

- $\text{VAP}_1 = \text{internal value alignment in humans}$: Making ourselves consistent.
  → difficult

- $\text{VAP}_2 = \text{internal value alignment in and between different groups of human agents}$: Social value alignment.
  → difficult
Making sets of value representations consistent

- $VAP_3 = \text{Stable individual-level human/machine consistency, after solving } (VAP_0 \land VAP_1) \\
  \rightarrow \text{very difficult, mainly because of } VAP_1$

- $VAP_4 = \text{Stable machine/society consistency, after solving } (VAP_0 \land VAP_1 \land VAP_3) \\
  \rightarrow \text{practically unsolvable}$

Asilomar AI Principle #10:

*Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.*
What is *intelligent* risk minimization?

→ The *causal root* of many major risk factors *really* are VAP$_1$ and (partially derived) VAP$_2$.

→ Resources should be allocated to (VAP$_1$ $\land$ VAP$_2$).

→ Resources include AI *itself*!
Variation: Scenario II

The empirical premise

5. The Superintelligence knows many things about us that we do not fully grasp or understand:
   ▪ It has a deeper understanding of the cognitive biases which evolution has implemented in our cognitive self-model and which hinder us in rational, evidence-based moral cognition.
   ▪ Empirically, it knows that the phenomenal states of all sentient beings which emerged on this planet—if viewed from an objective, impartial perspective—are much more frequently characterized by subjective qualities of suffering and frustrated preferences than these beings would ever be able to discover themselves, due to their evolutionarily developed mechanisms of self-deception.

6. It correctly concludes that human beings are unable to act in their own enlightened, best interest.
8. The superintelligence concludes that enlightenment is in the own best interest of all sentient beings on this planet.

9. Empirically, it knows that naturally evolved biological creatures are unable to realize this fact because of their firmly anchored existence bias.

10. It analyzes the human brain and realizes that a realistic potential for further mental evolution without suffering actually exists. → Premise 5 (b) is / can be made false.

11. The superintelligence decides to act benevolently.
The BAT-Scenario

**BAT**\textsubscript{Def.}: „Benevolent Artificial Transhumanism“

The emergence of a

- purely *ethically* motivated and
- genuinely altruistic
- form of transhumanism
- on epistemically superior postbiotic systems

is conceivable.
Can we control the process?

How Do We Align Artificial Intelligence with Human Values?

Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
Artificial Angel of Death or Bodhisattva-AI?

Practical and Ethical Implications of Artificial General Intelligence (AGI)

Thomas Metzinger
Gutenberg Research College
Philosophisches Seminar
Johannes Gutenberg-Universität Mainz
D-55099 Mainz

Frankfurt Institute for Advanced Studies
D-60438 Frankfurt am Main

metzinger@uni-mainz.de
http://www.philosophie.uni-mainz.de/metzinger/
Basic Copyright Notice & Disclaimer

©2018 This presentation is copyright protected. All rights reserved. You may download or print out a hard copy for your private or internal use. You are not permitted to create any modifications or derivatives of this presentation without the prior written permission of the copyright owner.

This presentation is for information purposes only and contains non-binding indications. Any opinions or views expressed are of the author and do not necessarily represent those of Swiss Re. Swiss Re makes no warranties or representations as to the accuracy, comprehensiveness, timeliness or suitability of this presentation for a particular purpose. Anyone shall at its own risk interpret and employ this presentation without relying on it in isolation. In no event will Swiss Re be liable for any loss or damages of any kind, including any direct, indirect or consequential damages, arising out of or in connection with the use of this presentation.