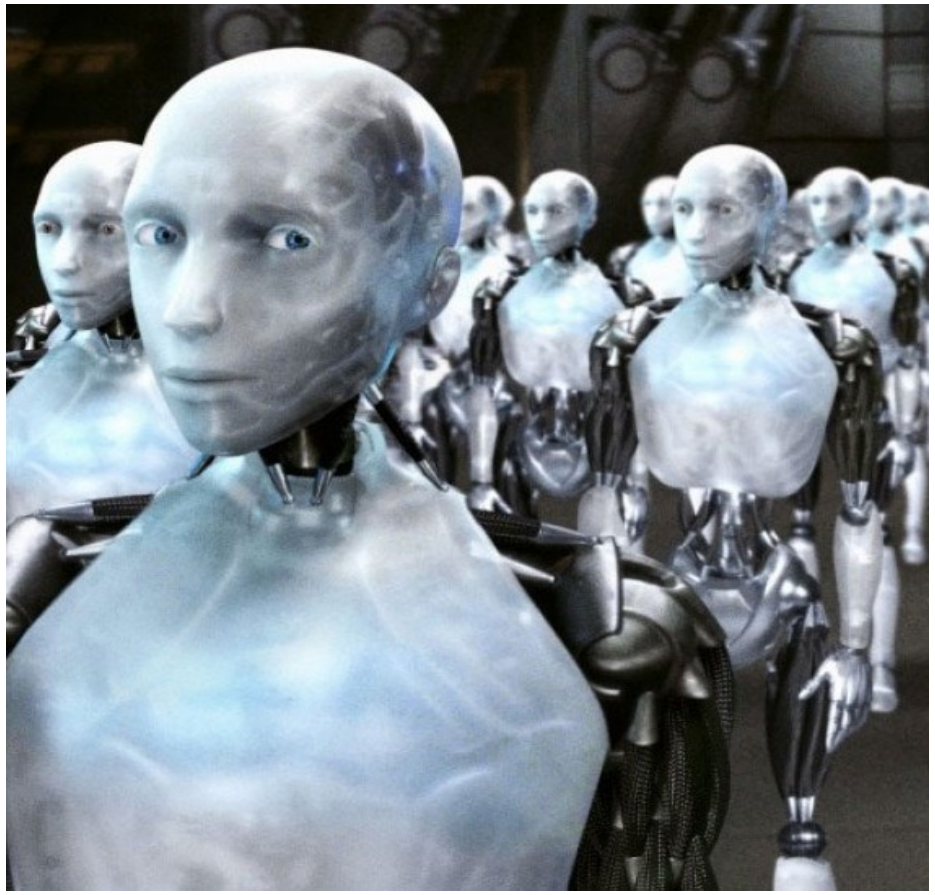
UPPSALA
UNIVERSITET

Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box?

Kathinka Evers

UPPSALA UNIVERSITET

Artificial Consciousness: Science Fiction?



The question whether a machine – a computer, a robot or any other form of artificial system – could be sentient is certainly entertaining, no end of science fiction deals with the question and sometimes very engagingly. But why is the question of artificial sentience (or “awareness”, or “consciousness”) raised in science and why invest public funding in this research? Is conscious AI at all possible, or even desirable?

UPPSALA UNIVERSITET

Three closely related questions

1. **Why strive to develop conscious artificial systems?**
   - Psychological & social driving forces
2. **Is artificial consciousness possible?**
   - Theoretical vs empirical possibilities
3. **Could artificial consciousness be recognised?**
   - The problems of gaming & commensurability

UPPSALA UNIVERSITET

1. Why strive to develop conscious machines?

The question why we would want to develop conscious machines, what the psychological and social driving forces are, is interesting to consider in historical perspectives on how the originally very limited circle of acknowledged minds in Western cultures took many centuries, even millenia, to expand.

UPPSALA UNIVERSITET

Society of “souls” in Western cultures

The question of consciousness was in Western cultures long raised in terms of possessing a "**soul**" understood as the immaterial aspect or essence of a human being, which partakes of divinity notably through its immortality.

This was a very exclusive society reserved for a limited number of people wanting to believe that they were “images of God”.

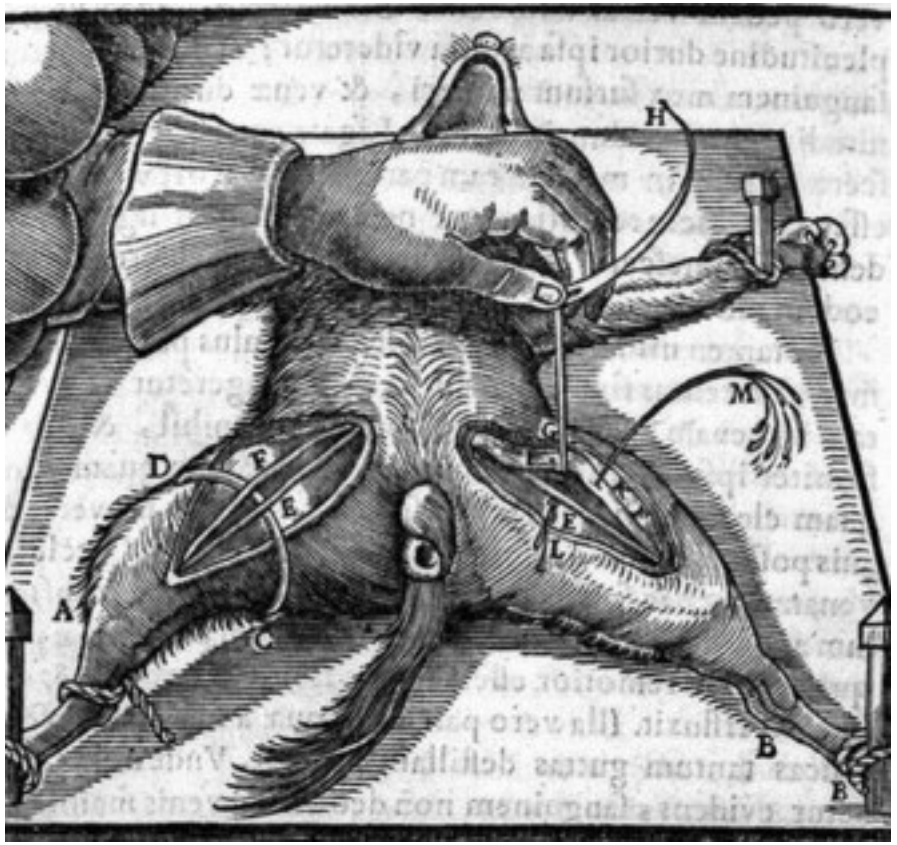
UPPSALA UNIVERSITET

Galen of Pergamum 129-216 C.E.
Animal suffering as inconsequential

The possession of an immortal soul was in Western cultures predominantly reserved for humans, whereas non-human animals were widely believed to exist solely to **serve human needs.**

UPPSALA
UNIVERSITET

Descartes 17thC
Vivisection of automatons?



UPPSALA UNIVERSITET

Mental hierarchies

Distinct concepts and related implications:

- Having or lacking a **soul** as a token of divinity and immortality
- Having or lacking **consciousness** (subjective experience) or **sentience** (valenced experience, feelings)
  - As an **either-or** situation
  - As a question of **grades,** levels or dimensions
- The **nature** of mental features in question: mental hierarchies
- The **ethical** implications drawn

Depending on era and cultural context, **humans were placed in different positions determined by ideologies, e.g., regarding gender, race, ethnicity or social class.**

UPPSALA
UNIVERSITET

1st Council of Nicaea y. 325: Do souls have gender?

UPPSALA UNIVERSITET

Subhuman souls?

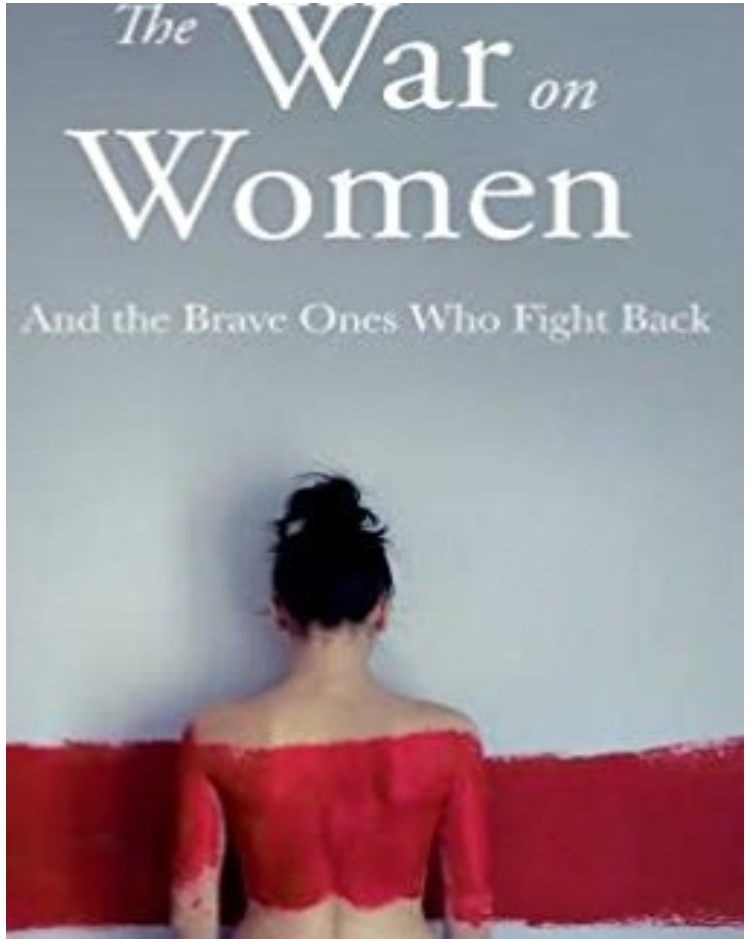
Déclaration des droits de l'homme et du citoyen (1789), Robespierre
Excluded women because they were not considered full-fledged humans

Déclaration des droits de la femme et de la citoyenne (1791),
Olympe de Gouge

Executed 1793

UPPSALA UNIVERSITET

Science as an ideological driving force



"The chief distinction in the intellectual powers of the two sexes is [shown] by man attaining to a higher eminence in whatever he takes up, than woman can attain…whether requiring deep thought, reason or imagination, or merely the use of the senses and hands."
Darwin: *The Descent of Man 1871*

Lloyd-Roberts: *The War on*

UPPSALA UNIVERSITET

The Indigenous Holocausts

Numerous cultures were eradicated by European colonialists in a series of genocides. Although the prime motivation was perhaps not the views that the Europeans held on consciousness or sentience, the **facility** of slaughtering populations or reducing humans to mere instruments is greatly enhanced by the view of them as lesser beings, or “human animals”, emotionally and intellectually.

UPPSALA
UNIVERSITET

The philosopher Herbert Spencer (*Social Statistics*, 1851) lauded imperialism for having exterminated sections of humanity that in their alleged inferiority “blocked the way for civilisation”.

UPPSALA UNIVERSITET

Human zoos. Ota Benga 1904, Bronx Zoo New York.

New York Times:

"A Bushman, one of a race that scientists do not rate high in the human scale."

Age 23 years. Height, 4 feet 11 inches. Weight 103 pound. Brought from the Kasai River, Congo Free State, South Central Africa, By D. Samuel P Verner. Exhibited each afternoon during September

UPPSALA UNIVERSITET

Why is this important?

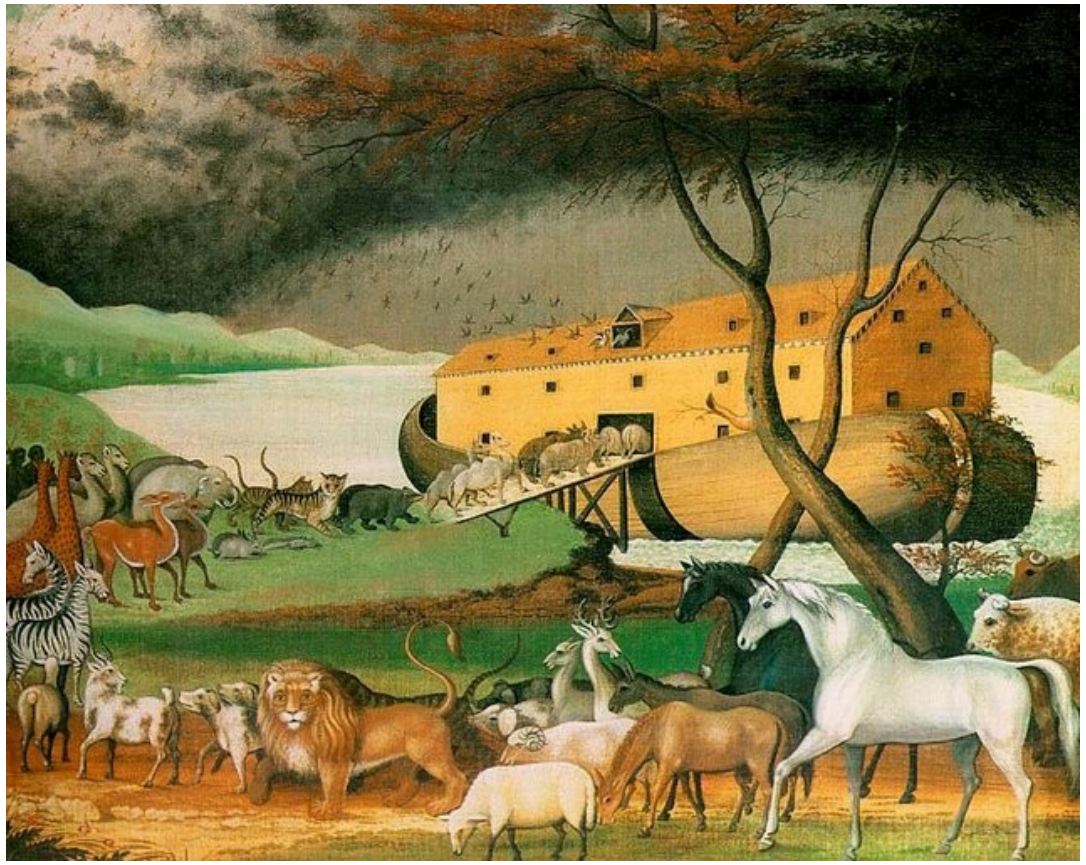
Whether we speak of biological or artificial entities, the question is not merely **whether** the soul/mind/sentience **is** there or not, but **how** it is there.

Science has in a very unscientific manner played a key role in establishing qualitative "mind scales" permeated with ideologies leading to unfathomable suffering.

It is difficult to assess either the presence or the qualities of the mind of another if your culture and the sciences that shape and are shaped by your culture dictate rejection.

UPPSALA
UNIVERSITET

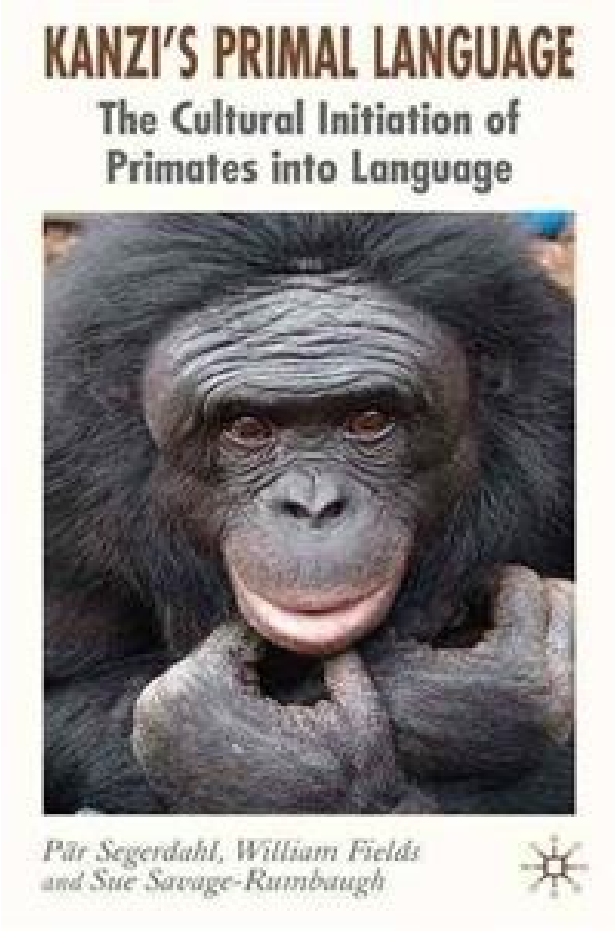
Expansion



After many centuries of ideological resistance, racial hierarchies are no longer in vogue in the sciences of mind, and the previously so passionately misogynistic “scientific” attempts to prove the inferiority of the female mind also seem to have lost at least some of their momentum.

UPPSALA
UNIVERSITET

Other Minds

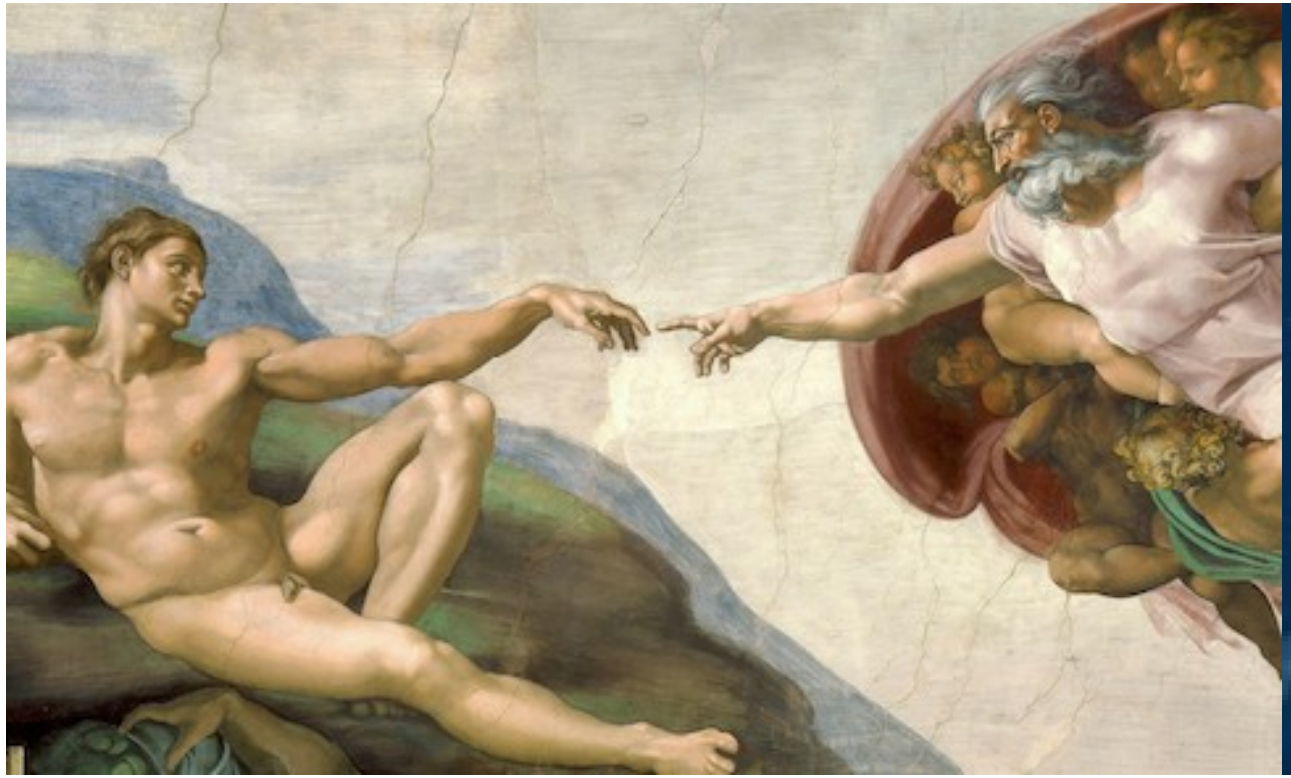


UPPSALA
UNIVERSITET

Artificial consciousness? Panpsychism?

UPPSALA UNIVERSITET

Psychological motivations: Intellectual openness or narcissism? The human as Creator



UPPSALA UNIVERSITET

Social motivations

Popular beliefs:

1. Consciousness could
   (a) enhance the capabilities of an artificial system, e.g., enable it to perform intentional moral decisions
   (b) which would increase possible benefits for humankind.
2. We need some kind of artificial awareness that some actions violate or risk undermining some human values, moral norms, etc:
   we need artificial awareness to align with human values.

UPPSALA UNIVERSITET

Machine benevolence and Isaac Asimov's "Three Laws of Robotics"

- Law One – "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
- Law Two – "A robot must obey orders given to it by human beings except where such orders would conflict with the First Law."
- Law Three – "A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law."
- Asimov later added the "Zeroth Law," above all the others – "A robot may not harm humanity, or, by inaction, allow humanity to come to harm."

UPPSALA UNIVERSITET

Utopia

- In one of Asimov's stories, robots are made to follow the laws, but they are given a specific meaning of "human." Prefiguring what now goes on in real-world ethnic cleansing campaigns, the robots only recognise people of a certain group as "human." **They follow the laws but still carry out genocide**.

- The most important reason for Asimov's Laws not being applied yet is **how robots are being used in our real world**. You don't arm a Reaper drone with a Hellfire missile or put a machine gun on a MAARS (Modular Advanced Armed Robotic System) not to cause humans to come to harm. That is the very point!

- Is it within my 2nd Amendment right to have a robot that bears arms?

Peter Singer: Isaac Asimov's Laws of Robotics are Wrong, The Brookings Institution, 2025

UPPSALA UNIVERSITET

A closer look at “the human creator”

- In human brains the capacity for other-oriented responses, such as **benevolence and sympathy, is pronouncedly selective and limited by spontaneous aggressive tendencies**. When sympathy and mutual aid is extended within a group, they are also (de facto) withheld from those that do not belong to this group. Interest in others is expressed towards **specific** groups and rarely extended **universally** to the human species, let alone to all sentient beings.
- **Human understanding of others does not entail compassion** but is frequently combined with emotional **dissociation from “the other”**.

UPPSALA UNIVERSITET

Innate tendencies & normative diversity

- Some **evaluative tendencies** may be innate features of the human species, for example: self-interest, control-orientation, dissociation, empathy, selective sympathy and xenophobia. By virtue of their historic prevalence, these features may be a part of our neurobiological identity and cultural imprints epigenetically stored in our brains.
- But there are few, if any, universal “human values”, or universally shared morality. To the contrary, **normative diversity** fundamentally characterises the human species – which no doubt forms a part of the cause of our tragic predicament.

Evers, K, Can we be epigenetically proactive? In T. Metzinger & J. M. Windt (2015) (Eds). *Open Mind: Philosophy and the mind sciences in the 21st century,* MIT Press, Cambridge, pp. 497-518.

UPPSALA UNIVERSITET

Some problematic human activities based on diverse “human values” (a selection)

- **Genocides** – the deliberate destruction of national, ethnic, racial, or religious groups – are presently ongoing, as in every preceding century (e.g., the slaughtered humans are regarded as “human animals”).
- **Femicides** – the deliberate murder of a woman by virtue of being a woman– are also a historical constant, presently committed every 10 minutes with reference to “values” (e.g., “honour”).
- **Ecocides**, e.g. the rapid annihilation of species today is estimated to be up to 10,000 times higher than the natural extinction rate (the rate of species extinctions that would occur if humans were not around).

UPPSALA
UNIVERSITET

In that light, conscious AI “alignment with human values” may well alarm more than it reassures!

Why do we believe that a violently destructive, xenophobic and misogynistic animal would create a universally benevolent machine?

Should we not rather hope that conscious AI would **not** align with either “human nature” or “human values”?

UPPSALA UNIVERSITET

Machine benevolence? Pandora's Box

- Humans are both **biologically and culturally predisposed** to act with violence towards **outgroup** individuals. Artificial systems are programmed with and have access to vast amounts of human-generated data, where **universal benevolence shines by its absence**.

- In that light, **the belief that a conscious machine created by humans would be engaged in universal human well-being appears at best unfounded.**

- Why should ”they” (machines) care about ”us” (humans)? And if they were to engage, why should they feel benevolence instead of malevolence towards outgroup individuals (that would be far more in line with the mind of their programmer, their human “creator”)?

In view of how selectively and xenophobically conscious intelligence operates in humans, taking machine benevolence towards humans for granted appears dangerously naïve.

UPPSALA UNIVERSITET

Machine welfare: an unlikely scenario

Reversing the perspective: machine consciousness and sentience introduces the issue of machine welfare. Seeing how humans treat other animals and how humans treat other humans, there is ample reason to doubt that machines would face a happy destiny if we, whether intentionally or inadvertently, were to construct machines capable of reason and emotion. And, as in the case of, e.g., vivisected animals, or enslaved and intellectually stunted humans, their suffering might long go unacknowledged and even undetected.

Irresponsibility is added to naïvety in the narcissistic dream to create conscious machines.

Is precaution needed?

UPPSALA UNIVERSITET

2. Is artificial consciousness possible?
Epoché

- Presently, consciousness is only known to exist in living things. That is a fact about our knowledge that does not logically exclude artificial consciousness.
- Conscious AI is assumed to be theoretically possible within certain theoretical frameworks.
- For now, no independent empirical evidence is available.

In that situation, we can neither logically exclude nor affirm the possible existence, or future existence, of artificial consciousness in the real world.

UPPSALA UNIVERSITET

Functionalism: irrelevance of substrates

‘Conscious processing may be implemented in exactly the same way in different physical substrates, whether biological or artificial. If the system functions in the right way, it can be conscious, whatever it is made of, for the substrate and its architecture are irrelevant.’

IF an animal brain could be emulated neuron-by-neuron and the emulation were put in control of a robot, **then**, if the same pain markers that were accepted to indicate pain in the animal were present in the robot, we should in the name of consistency, **all other things being equal**, draw the conclusion that the robot might also feel pain.

Birch & Andrews, Intellectica

UPPSALA UNIVERSITET

The relevance of life to sentience

Consistency dictates that if two entities, x and y, share the same feature, f, and we draw a conclusion (e.g., the presence of sentience) about x with reference to f, then, all other things being equal, we should draw the same conclusion about y with reference to f. **But are all other things equal in this case?**

- One substrate is **alive**, the other is not, and this introduces a potentially huge and, epistemically as well as morally, relevant difference between the two cases.
- We cannot simply assume the **contingency of life for sentience** and take the possibility of non-living, e.g., artificial sentience for granted.
- A possible reply is that **sentience entails life** so a sentient robot would be alive, thus reducing the relevant difference between the two substrates. The relation between life and sentience, as well as each of those concepts would still stand in need of further clarification. Likewise, the general question which features (if any) are **essential**, or indispensable, for sentience and which are contingent.

Cf. Discussion in Intellectica (2024/2: 81): Birch & Andrews, Commentary by Evers

UPPSALA UNIVERSITET

3. Could artificial consciousness be detected?

If consciousness were to exist in an entity which by its constitutive nature is materially different from living, biological brains, would it be similar to ours?

By what reasoning may we justify an answer? If it is not similar, how might this affect our abilities to (a) **detect** it, (b) **understand** or gain knowledge of it, and (c) **communicate** with it, provided some success in (a) and (b)?

UPPSALA UNIVERSITET

The gaming problem

- Artificial systems use human-generated training data to mimic human behaviours, which, if successful, may persuade human users of their sentience. Here we are in **the realm of psychology** rather than logic:
- **Logically**, mimicking human behaviours in artificial systems have **no evidential value** whatsoever.
- The gaming problem does not occur to equal extent with animals, since they have evolved without using human-generated training data to mimic human behaviours.

UPPSALA UNIVERSITET

Switching focus from similarity to difference

"Most AI works very differently from a biological brain. It isn't the same functional organisation in a new substrate; it's a totally different functional organisation." (Birch & Andrews, 2024)

"In other words: it is a totally different functional organisation in a totally different substrate that has a totally different architecture. Quite possibly, its sentience – if present – would also be totally different, and it is via those **differences** that it might best be detected." (Evers 2024)

UPPSALA
UNIVERSITET

If, say, an artificial system shows signs of enjoying music without being programmed to do so and plays what humans might like, we would be faced with the gaming problem, whereas if it plays something humanly abhorrent (for example, mixing simultaneously three pieces combining Bach, rap and lullabies speeding it all up to play a hundred times faster in multiple repetition), we might still be faced with the gaming problem, but in a more interesting and thought-provoking way.

UPPSALA UNIVERSITET

The problem of commensurability

- A challenge in detecting differences is that we cannot think entirely beyond our own perspective, we are imprisoned by the limits imposed by our bodies, so if the differences are sufficiently deep, we cannot detect them. There must also be some similarities to justify the application of the same concept to distinct phenomena, so a total difference might by necessity remain beyond our reach.
- If animals and artificial systems are ”totally different” substantially, structurally and functionally, animal and artificial sentience (assuming that the latter concept makes sense) might also be totally different and, therefore, at least to some extent, **incommensurable** which would pose a formidable obstacle for detecting, let alone understanding, a sentient machine.

Evers, K., Farisco, M., Pennartz, C.M.A. (2024) Assessing the commensurability of theories of consciousness: On the usefulness of common denominators in differentiating, integrating and testing hypotheses. Consciousness and Cognition, 119. https://doi.org/10.1016/j.concog.2024.103668

UPPSALA UNIVERSITET

Strategies for finding bench-marks of conscious AI

- **Theory-based** strategy: starting from selected theories of consciousness in order to infer relevant indicators (Butlin, Long et al. 2023)
- **Life-based** strategy: consciousness necessarily connects with biological life (Searle 2007, Godfrey-Smith 2016, Seth 2024)
- **Brain-based** strategy: the brain, its evolution, and its correlation with consciousness are benchmark for artificial consciousness (Aru, Larkum et al. 2023; Farisco, Evers & Changeux, 2024)
- **Consciousness-based** strategy: searching for other forms of biological consciousness than the human, in order to identify what (if anything) is indispensable to consciousness and what is contingent (Birch and Andrews 2024)
- **Heuristic approach**: elaborating a list of indicators of consciousness in artificial systems (Pennartz, Farisco & Evers 2019) and related tests for artificial consciousness (Elamrani & Yampolskly 2019; Bayne, Seth et al, 2024)

(Farisco & Evers, Is it possible to identify phenomenal consciousness in artificial systems in the light of the gaming problem?, In preparation.)

UPPSALA
UNIVERSITET

There are many interesting attempts to search for bench-marks of artificial consciousness. Yet the results of this research might well in the end tell us more about human beings than about the artificial systems studied.

UPPSALA UNIVERSITET

Grounds for scepticism?

1. Conscious AI may be possible in some theoretical frameworks, but no empirical evidence suggests that it might be possible in reality.

2. Due to deep differences in substrate, architecture and functions, artificial consciousness might be fundamentally different from human and animal consciousness (if it were to exist).

3. Due to the problems of gaming and incommensurability, artificial consciousness might be impossible to detect, let alone understand.

4. Because of how human nature has been expressed throughout our history, developing conscious machines (possibly with for humans undetectable and incomprehensible minds) is a monumentally dangerous idea.

UPPSALA UNIVERSITET

Conclusion

Conscious AI created by humans would likely be

(a) monumentally dangerous
(b) very different from human consciousness and thus possibly
    a. undetectable
    b. incomprehensible
    c. incommunicable

(b) increases (a) and cuts both ways, if AI has valenced experience

UPPSALA
UNIVERSITET

Thank you!