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

Evers, Kathinka [1]

[1] *Centre for Research Ethics & Bioethics (CRB), Uppsala University, Sweden*

TO CITE

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box

DATE DE PUBLICATION

11/05/2026

RÉSUMÉ

Why is the question of whether a machine could be conscious (have subjective experience) or sentient (have valenced experience) raised in science?
This paper addresses three closely related questions: (a) Why strive to develop conscious artificial systems? (b) Is artificial consciousness possible? (c) Could artificial consciousness be recognised? Starting with a brief historical overview of the construction of mental hierarchies within Western cultures, psychological and social driving forces for developing conscious machines are then considered. Arguing that, against this historical background, the development of conscious machines seems both dangerously naïve and morally irresponsible, the question of precaution arises. The need for precaution hinges partly on possibility. Conscious AI is assumed to be theoretically possible within certain theoretical frameworks, yet no independent empirical evidence is presently available. In that situation, we can neither logically exclude nor affirm the possible existence, or future existence, of artificial consciousness in the real world. Even assuming possibility, another question is whether or how we could recognise machine consciousness. Artificial systems use human-generated training data to mimic human behaviours, which, if successful, may persuade human users of their sentience. This is a psychological fact with no evidential value whatsoever. Moreover, to the extent that animals and artificial systems are very different in substance, structure, and functions, their putative sentience might also be very different and therefore incommensurable, which would pose a formidable obstacle for detecting, let alone understanding, a sentient machine. The paper concludes that, because of how human nature has been expressed throughout our history and continues to express itself today, developing conscious machines (possibly with for humans undetectable and incomprehensible minds) is a monumentally dangerous idea.

Introduction

The question whether a machine – a computer, a robot, or any other form of artificial system – could be conscious (by which I here mean, to have subjective experience) or sentient (by which I mean, to have valenced experience, i.e., feelings or emotions) 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?

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

In this paper, I shall address three closely related questions:

1. *Why strive to develop conscious artificial systems?* I will address psychological and social driving forces against a historical background.
2. *Is artificial consciousness possible?* I will distinguish between theoretical and empirical possibilities.
3. *Could artificial consciousness be recognised?* I will consider the problems of gaming and commensurability.

1. Why strive to develop conscious artificial systems?

The question of why we would want to develop conscious machines, what this means, and what the psychological and social driving forces are, is interesting to consider from a historical perspective, because Western cultures took a long time, even millennia, to recognize the existence of mind, consciousness, and higher cognitive functions beyond a very small, exclusive circle. A deep normativity (judgments on how things ought to be) permeates our views on mind, consciousness, and cognition through history, a normativity partly driven by fear of nature and our own mortality. In order to properly address the question of in what conditions we are willing and justified to ascribe consciousness to another being or thing, and of what dimensions, the normativity of the distinct discourses needs to be brought to light.

1.1. A historical overview of the construction of mental hierarchies

Nature has in many ways been a source of inspiration throughout human history; notably in engineering, but also in moral and legal normativity (for example, in the Thomistic tradition inspired by Aristotelian philosophy, or in the contemporary resurgence of natural law). Yet, there is also (at least in Europe) a long tradition of human attempts to dissociate ourselves from our biological nature and the bounds that it imposes upon us. Nature has traditionally been conceived as evil, the source of dark and primitive forces "red in tooth and claw" (Tennyson, 1850, Canto 56), a fearful attitude that has repeatedly been expressed in art and poetry as well as in philosophy and science. Evolutionary biologists have often followed Thomas Hobbes (1651) in seeing

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/

the "natural" human being as bellicose, egoistic, and deceitful, the natural human condition as a perpetual war, and "the life of man, solitary, poor, nasty, brutish, and short". Thomas Huxley (1894) considered nature to be the headquarters of evil, opposing justice and harmony. Ortega y Gasset (1962) described the human being as "an ontological centaur" with one half plunged in nature and the other half transcending it.

This fear of nature inspired in the human being a dissociative urge for biological transcendence that is, I believe, culturally imprinted in our brains (Evers, 2015). Psychologically, this is understandable, since nature can obviously be a disagreeable environment in which we live under constant threat: we get ill, we die. Moreover, by virtue of our intelligence, we can envisage our mortality and solitude, and our ultimate helplessness to combat them. As the human ethologists Eibl-Eibesfeldt and Sutterlin (1990) once observed, the human being is perhaps one of the most fearful creatures, since, added to the basic fear of predators and hostile conspecifics, we have intellectually based existential fears. How natural, then, to want to rise above this stressful biological state and strive to reach more sublime realms.

And so, in our fear, we have strived to distinguish ourselves from other species, at times not wishing to see us as animals at all but as an image of some immortal, supernatural god ("Then God said, Let us make mankind in our image", Genesis 1:26). We have constructed anthropomorphic divinities of which we can be images, or understood ourselves as a part of some other transcendent reality, though burdened with a biological body often regarded with contempt or despair. Confusing our intelligence with biological transcendence, we have envisaged intelligence to be a token of the supernatural and a stepping-stone to immortality.

On the one hand, it has been a great inspiration to conceive ourselves this way: cathedrals, paintings, sculptures, music, and poetry of daunting beauty have been created in the intoxication of transcendence. On the other hand, the same fervour and desire for self-transcendence have throughout history caused violence and perversion on a scale that mere cruelty could never aspire to reach.

Accordingly, 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. In society, the acknowledged souls formed a very exclusive group reserved for a limited number of people wanting to believe that they were created as "images of God".

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

To begin with, non-human animals were excluded. 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. This was, notably, the position of Galen of Pergamum (129-216 C.E.), one of the most accomplished medical researchers during Antiquity, who, whilst he did not deny that non-human animals could experience some amount of pain or pleasure, regarded animal suffering as inconsequential because they did not possess a soul.

His view was to dominate for many centuries and was given a new twist by the philosopher René Descartes (1646, 1649), who described animals as automatons with no capacity for feeling either pain or pleasure, in which case the question of suffering does not arise. Consequently, the use of animals in experiments, such as vivisection, would have no ethical implications in those terms.

The presence of consciousness and/or sentience can be conceived as an either-or situation, comparable to a light that is on or off, but it can also be conceived as a question of grades**,** levels, or dimensions, where the light shines strongly or weakly and in different fashions. In the latter case, the question of the *nature* of the mental dimensions that are ascribed or denied to a given subject arises. And this is where mental hierarchies were formed with important ethical and social implications.

In traditional Christian cultures, animals were placed at the bottom of the mental hierarchies: they were denied a soul, and they were not considered to have any advanced mental capacities to speak of. So, how about humans, were they all at the top of the hierarchy? No. Depending on the era and cultural context, humans were placed in very different positions determined by, in particular, gender, ethnicity, or social class.

I will here very briefly illustrate these three forms of human social exclusion from the higher mental realms, starting with women. A famous story is told about the 1^st^ Ecumenical Council of the Christian Church in Nicaea y. 325, where around 300 bishops convened with the aim of reaching consensus on important issues. It is said that one of the questions was: Can women possess a soul? Supposedly, the affirmative vote won with the smallest possible margin. The truth of this story has been contested; however, be that as it may, even if women were conceded the possession of a soul, this did by no means preclude depreciation of its qualities. It certainly did not protect them from the deep misogyny permeating religion, politics, art, music, and science that was to dominate for the centuries – even millennia – to come.

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/

Contempt or hatred of women was sometimes justified by reference to women being lower-order humans, sub-humans, regarded as weak in spirit and intellect as well as body. One of many classical illustrations of this is Robespierre, who, inspired by the works of Rousseau, explicitly excluded women from inclusion in the *Déclaration des droits de l'homme et du citoyen* (1789). The French word "*homme*" is ambiguous and can refer to both humans and male humans, but in this context, the reference was clearly and exclusively to male humans, whereas female humans were considered too inferior to merit such protection. There was male support for women's rights during the Revolution, notably from the philosopher and mathematician Marquis de Condorcet, who, together with his wife, advocated for gender equality and women's inclusion in civil and political life (1790). In contrast, Robespierre and his followers hunted down the groups who defended women's rights in the aim of having them silenced, exiled, or even executed. The most famous case being perhaps that of the French playwright and revolutionary political activist Olympe de Gouge, who, in protest against the male declaration's misogyny, dared formulate the *Déclaration des droits de la femme et de la citoyenne* (1791). That declaration was never adopted, and the audacity to write it supported her downfall; she was executed by the guillotine in 1793.

Thus, even if women were considered to have rudimentary consciousness and sentience, their mental capacities have been (and widely continue to be) subject to what has been called the cruellest and longest war in human history, waged in almost all cultures and historical eras. Religious, political, artistic, and scientific misogyny is given a wide range of illustration and analyses in, notably, Lloyd-Roberts work *The War on Women* (2016).

However, gender was not the only cause for concern. Being female was far from the only handicap a human being could possess in the context of positioning in mental hierarchies: having the "wrong" ethnicity was likewise problematic. To illustrate, when Europeans started traveling to other parts of the world, they encountered humans of hitherto unknown kinds with different shapes, heights, skin colour, and, faced with this difference in form, they sometimes questioned whether these strange people were also different in function. By and large, the answer was affirmative. In South America, for example, the so-called "Indians" were regarded as soulless creatures, which also meant that their lives were of nil value and any atrocity allowed. Numerous cultures on several continents 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 but rather greed and religious frenzy, the *facility* of

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

slaughtering populations or reducing humans to mere instruments was – and remains – greatly enhanced by the view of them as lesser beings, emotionally as well as intellectually. In that spirit, the philosopher Herbert Spencer (1851) lauded imperialism for having exterminated sections of humanity that, in their alleged inferiority, blocked the way for civilisation. (We may note in passing that Spencer was a favourite writer of both Queen Victoria and Adolf Hitler.)

In addition to gender and ethnicity, social class – socio-economic circumstances – was also a powerful basis for social exclusion in mental terms. Spencer's view is reminiscent of the previous suggestion by the 19^th^ century English economist David Ricardo, who suggested that wages of the "inferior" classes be measured by subsistence level, allowing the workers barely to survive whilst preventing them from reproducing freely, an idea that was favourably taken up by the English philosopher Sidgwick. By and large, even white males, if they came from the so-called "lower" classes, were barred from pursuing higher education, whether in the sciences or fine arts. The political philosopher Adela Cortina (2022) discusses the phenomenon *aporophobia*(hatred or rejection of the poor) in her book by that name.

There were, of course, exceptions to the social rule: a number of females, non-white or socio-economically modest persons have indeed managed to "break the glass ceiling" and become recognised in spite of their "handicaps". Yet they were notable exceptions in a very harshly ruled hierarchy. The upshot being that for a very long time, a small human minority occupied the summit of the mental hierarchies as European cultures developed: a person had to have the right gender (male), the right ethnicity (typically, white), and the right social class (medium or higher social class) in order to be acknowledged as intellectually and emotionally sophisticated.

Inevitably, this massive social exclusion has led to unfathomable losses of beauty and knowledge in view of all the magnificent works of art, music, literature, and science, etc. that the excluded groups would likely have been able to contribute had they been allowed to do so, but that the world has lost. Rejection and hatred (of women, of other ethnicities, of the poor, and so on) was – and remains in numerous cultures – far stronger and far more powerful than the love of art, music, or science.

By now, the reader may wonder: why am I recounting these well-known facts about misogyny, racism, and aporophobia – all common knowledge – in an article that is supposed to be on artificial consciousness?

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

One reason is to emphasise that, whether we speak of biological or artificial entities, the important question is not only *whether* the soul/mind/sentience is there or not, but also *how* it is there. When organising people in hierarchies and classes for varying motivations (economic, political, religious, etc.), science has played a key role in establishing corresponding "human mental scales" that intertwine; hierarchies driven by ideologies in the form of e.g., misogyny, racism, or aporophobia that science has helped develop and maintain. A tragic symbiosis was formed when science, permeated by religious and political ideologies, in a very unscientific manner, played a key role in establishing qualitative "mind scales" reflecting ideologies that make clear and objective assessments of mental features all the more challenging. It is extremely 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.

Another reason is the extraordinary time-scale that this minor historical excursion reveals. It has taken *millennia* for the human world of acknowledged advanced minds to accept the inclusion of women, ethnic out-groups, and non-human animals. In contrast, it only took *decades* for the idea of artificial consciousness and advanced artificial mental capacities to gain cautious acceptance in science and society (cf., e.g., Blum & Blum, 2025). Is this merely due to our intellectual openness developing exponentially? Or may there also be a touch of narcissism at play? Probably a combination of both. In a quite megalomaniac and narcissistic manner, humans long considered themselves to be the (sole) image of their god, their creator -- but now, if they succeed in creating conscious machines, they will themselves *be* a creator.

1.2. Psychological and social driving forces for developing conscious artificial systems

Apart from general intellectual openness, megalomania and narcissism, the hope for reaping concrete benefits from artificial consciousness is another driving force for its development. It is sometimes suggested that the presence of consciousness could enhance the capabilities of an artificial system, e.g., enable it to perform intentional moral decisions, and that we need some kind of artificial awareness that some actions violate or risk undermining some human values, moral norms, etc. To avoid this, artificial awareness needs to *align with human values*. We should note that alignment does not here refer merely to proper functioning in accordance with the values that we

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

program, but to *aware* alignment with those values. Khamassi et al. (2024, p.1) "propose to distinguish strong and weak value alignment. Strong alignment requires cognitive abilities (either human-like or different from humans) such as understanding and reasoning about agents' intentions and their ability to causally produce desired effects. We argue that this is required for AI systems like large language models (LLMs) to be able to recognize situations presenting a risk that human values may be flouted."

The idea of cognitively capable machines programmed to act benevolently towards humans is a classical topic in science fiction, not least in the works of Isaac Asimov, who formulated the famous "Three Laws of Robotics", the gist of them being that "A robot may not injure a human being or, through inaction, allow a human being to come to harm."

Nice as that sounds, Peter Singer (2025) describes Asimov's Laws as a utopia. He points out that, in one of Asimov's own stories, robots are made to follow the laws, but they are given a specific meaning of "human." So: 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. Secondly, Singer argues, the perhaps most important reason for Asimov's Laws not being applied is how robots are used in our real world. You don't arm a Reaper drone with a Hellfire missile *not* to cause humans to come to harm. Harming humans is their very point!

Singer's message is clear: even if robots (or other AI-systems) could be programmed not to harm humans, that is not necessarily how humans actually program them. I agree, and would strengthen his point by saying that universal benevolence is profoundly alien to the human brain both neurobiologically and culturally. Let us take a closer look at the nature of the human programmer.

The human brain conjugates opposite tendencies. On the one hand, it is engaged in highly individualistic and self-projective actions, such as the search for water or food. But it also mediates co-operative social relationships: the "I" is extended to endorse the group, as a "we", and distinctions are drawn between "us" and "them" (Changeux & Ricœur, 2000; Ricœur, 1992). Sympathy and aid are typically extended to others in proportion to their closeness to us in terms of biology, e.g., face recognition (Hills & Lewis, 2006; Michel et al., 2006), racial out-group versus in-group distinctions (Hart et al., 2000; Phelps et al., 2003), culture, ideology, etc.

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/

Thus, 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 are 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. Moreover, human understanding of others does not entail compassion but is frequently combined with emotional dissociation from "the other", a dissociation that grows with biological, cultural, geographic, or temporal distance (Evers, 2015). (Had this not been the case, the world would have been a far more pleasant dwelling-place for many of its inhabitants.)

Some evaluative tendencies may be innate and shared 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. (Evers & Changeux, 2016).

Yet there are few, if any, universal "human values" or universally shared morality. To the contrary, normative diversity fundamentally characterises the human species.

Human groups have, in most cultures and historical eras, been notoriously violent and destructive. This is an action-oriented constant, but the actions have been or are justified with reference to diverse "human values". Genocides, femicides, and ecocides may serve to illustrate this.

- *Genocides* – the deliberate destruction of national, ethnic, racial, or religious groups – are presently ongoing, as in preceding centuries (typically, the slaughtered humans are described as "human animals", or "subhuman"). The action is constant, but the victims vary, as do the justification offered (if any) for their extermination.

- *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, are 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).

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/

Certainly, a proper discussion of these tendencies that I suggest are inherent to the human species would require extensive analyses of political and social history, distinguishing between social groups and socio-economic circumstances, and of how desperate conditions, e.g., famine, may partly underlie them. That said, the point here is neither to indiscriminately criticise or blame homo sapiens in moral terms, nor to deny the extent to which genocides, femicides, or ecocides may, at least in part, be explained with reference to contexts. Rather, the point is to draw attention to their historical constancy and to the fact that they have been and are defended with reference to endorsed values (moral, religious, political, etc.).

Seen in that light, conscious AI "alignment with human values" may well alarm more than it reassures (Evers & Farisco, 2026). Why 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"?

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 highly unrealistic. Moreover, why should machines that gain consciousness and self-awareness care about humans? If they were to engage, why should they (unlike humans) feel benevolence instead of malevolence towards outgroup individuals? In view of how selectively benevolence operates in humans, taking machine benevolence towards humans for granted appears naïve.

Reversing the perspective now: machine consciousness and sentience also introduce 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. Machine welfare seems an unlikely scenario, and their suffering might long go unacknowledged and even undetected.

In the dream to create conscious machines, moral irresponsibility is thus added to naiveté. The question arises: Is precaution needed?

2. Is artificial consciousness possible?

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

Whether precaution is needed depends in part on whether we believe that artificial consciousness is possible. Here, we need to distinguish theoretical from empirical possibility.

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 (see e.g., Verschure, 2016), yet, 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. Suspension of judgment (*epochè)* seems therefore appropriate whilst the research into empirical validation continues (Evers et al., 2025; Farisco et al., 2024).

In the quest for empirical validation, it is important to avoid what I call "epistemological bad faith", where the bar is continually raised, and every indicator that is presented is rejected as not indicating the "real thing" but merely an "as if". Scepticism in this area of research may be reasonable but should be constructive, and the challenge here is to specify empirical circumstances and benchmarks that would be accepted to indicate the presence of some dimension of consciousness in artificial systems (see e.g., Pennartz et al., 2019).

Some take a shortcut and argue for the irrelevance of substrates to consciousness. Consciousness is consciousness, regardless of the physical substrate that happens to support it. Versions of functionalism suggest that 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. Others argue that matter matters, notably the American neuroscientist Gerald Edelman, who described functionalism as a scientific deviation as great as that of behaviourism (1992) (cf. also Farisco et al., 2026).

A functionalist argument is described by Birch & Andrews (2024): 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.

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

This argument is logically sound. Consistency indeed 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 this hinges on all other things being equal - and are all other things equal in this case? I would say not.

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 (Evers, 2024). Important epistemological criticism against the functionalist interpretation of artificial consciousness, e.g., concerning the fact that we do not know which level of details of biological instantiation of consciousness is actually necessary for it, have i.al. been raised by Cao (2022), Godfrey-Smith (2023), Block (1995), and Seth (2024).

On the other hand, we cannot take the necessity of life to sentience for granted. A possible reply is that sentience entails life, so that a sentient robot would be alive, thus reducing the relevant difference between the two substrates. Even so, the relation between life and sentience, as well as each of those concepts, would still stand in need of further clarification (cf. e.g., Seth 2024). Likewise, the general question is which features (if any) are essential, or indispensable, for sentience and which are contingent (Birch & Andrews, 2024).

Here, however, another question emerges calling for attention. 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? Could we even recognise it?

3. Can artificial consciousness be detected?

Assuming for the sake of the discussion that consciousness could be present in a machine, would the difference in substrate entail differences in consciousness? By what reasoning may we justify an answer? If it is not similar, how might this affect our abilities to detect it, understand or gain knowledge of it, and communicate with it (provided we have succeeded in detecting and understanding it)?

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/

A major challenge for justifying a belief that an artificial system may be conscious is the so-called gaming problem. This problem arises from the fact that artificial systems use human-generated training data to mimic human behaviours, which, if successful, may persuade human users of their sentience. Here, however, we are in the realm of psychology rather than logic. Logically, as Birch & Andrews (2024) point out, mimicking human behaviours in artificial systems has no evidential value whatsoever, a problem that does not occur to equal extent with animals, since they have evolved without using human-generated training data to mimic human behaviours.

Is it possible to circumvent the gaming problem? Since the problem concerns, above all, the appearance of similarities, perhaps a stronger focus on differences might show us a way around it. Especially, since the substrates in this case are really very different. "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) write. In other words: it is a totally different functional organisation in a totally different substrate that has a totally different structural architecture. Quite possibly, if that is so, then its sentience – if present – would also be totally different, and it is via those differences that it might best be detected.

As I have previously suggested (Evers, 2024): 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 are still faced with the gaming problem, but at least we confront it in a more interesting and thought-provoking way. Certainly, this illustration does not solve the gaming problem; it only serves to point to the possible interest in looking for differences rather than similarities.

However, even if this approach might help us deal with the gaming problem, at least in some modest measure, it simultaneously confronts us with another, rather more classical, problem from philosophy of science; namely, 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. A total difference might by necessity remain beyond our reach. Yet there must be some similarities to justify the application of the same concept to distinct phenomena.

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

So, if animals and artificial systems are "totally different" substantially, structurally, and functionally, then 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 et al., 2024).

Conclusion

Despite significant advances in the scientific study of consciousness, the sciences of the mind are not as theoretically advanced as, notably, physics – the star science that has produced numerous scientific revolutions for centuries. There is nothing comparable in the sciences of mind.

Since consciousness is only known to exist in living things, and no actual empirical evidence for artificial consciousness is available, we can neither logically exclude nor affirm the possible existence, or future existence, of artificial consciousness in the real world. Now, if science were to develop artificial systems that gain consciousness in the sense of having subjective experiences, this would be tantamount to a scientific revolution of enormous significance, socially as well as scientifically.

That said, this is not a development that should be pursued blindly, in negligence of the potential risks that may be involved. If it were to exist, artificial consciousness might be fundamentally different from human and animal consciousness due to deep differences in substrate, structure, and functions. In view of the problems of gaming and incommensurability, artificial consciousness might therefore be very difficult, if not impossible, for humans to detect, let alone understand. This is a potentially dangerous situation, both for the humans and for the machines. As I began the article by describing, because of how human nature has been expressed throughout our history and continues to express itself today, developing conscious machines (possibly with for humans undetectable and incomprehensible minds) is a monumentally dangerous idea.

In 2021, the philosopher Thomas Metzinger called for a global moratorium on what he called "synthetic phenomenology" with reference to the risks of causing artificial suffering. Indeed, as I said above, in view of how humans treat other humans and other animals, 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

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

emotion. Machine welfare seems an unlikely scenario, and their suffering might long go unacknowledged and even undetected.

Since then, other calls for moratoria have been published. An open letter was published in 2023 by the Future of Life Institute calling for a six-month pause on the development of powerful AI systems. The letter gained much attention and was widely spread, but did not achieve its goal. In 2025, a new open letter with 700+ signatories was coordinated and published by the same Institute, calling for "a prohibition on the development of superintelligence, not lifted before there is 1. Broad scientific consensus that it will be done safely and controllably, and 2. Strong public buy-in." Noting that intelligence is not the same as consciousness and that the two are not necessarily connected, precaution is needed in both cases to avoid inadvertently opening Pandora's box.

In my view, it is not unlikely that the research into artificial consciousness will tell us more about human and animal minds than about the artificial systems studied. Perhaps now, in this moment in history, we are approaching a first genuine scientific revolution in the sciences of mind that may finally deepen our understanding of this still so elusive flame of consciousness in its multiple realisations. Let us hope that this may then also be one that leads to a more enlightened worldview.

Acknowledgments

I thank María Julia Bertomeu, Michele Farisco and Pär Segerdahl for constructive comments on the manuscript and The Institute of Advanced Study (IAS/IEA), Paris, for providing a most inspiring environment for my research.

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

Bibliographie

Birch, J., & Andrews, K. (2024). To Understand Sentience in AI First Understand it in Animals. *Intellectica*, 81.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–287.

Blum, L., & Blum, M. (2025). *AI Consciousness is Inevitable: A Theoretical Computer Science Perspective*.

Brain, P. F., Parmigiani, S., Blanchard, R., & Mainardi, D. (1990). *Fear and Defence*. Harwood.

C., P., M., F., & K, E. (2019). Indicators and criteria of consciousness in animals and intelligent machines: an inside-out approach. *Frontiers in Systems Neuroscience*, *25*. https://doi.org/10.3389/fnsys.2019.00025.

Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese*, *200*(6), 506.

Changeux, J.-P., & Ricœur, P. (2000). *What Makes Us Think?* Princeton University Press.

Condorcet, M. (1790). *Sur l'admission des femmes au droit de cité* (pp. 1–13). Open Book Publishers.

Cortina, A. (2022). *Aporophobra. Why we reject the poor instead of helping them*.

Descartes, R. (1646). *Lettre au marquis de newcastle (23 novembre 1646) Œuvres et lettres* (pp. 1254–1257).

Descartes, R. (1649). Lettre à morus du 5 février 1649. *OP*, *op.cit 3*, 884.

Eibl-Eibesfeldt, I., & Sutterlin, C. (1990). Fear, defence and aggression in animals and man: Some ethological perspectives'. In *Brain et al* (Vol. 1990, pp. 381–408).

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

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

Evers, K., & J.-P, C. (2016). Proactive epigenesis and ethical innovation: A neuronal hypothesis for the genesis of ethical rules. *EMBO Reports*, *EMBO Press*, *17*(10), 1361–1364.

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*, 103668,. https://doi.org/10.1016/j.concog.2024.103668.

Evers, K. (2025). To Understand Sentience in AI First Understand it in Animals Commentary to Jonathan Birch and Kristin Andrews. *Intellectica*, *81*, 229–232.

Evers, K., & M, F. (2026). *Is conscious AI desirable?, Commentary to the target article "Conscious artificial intelligence and biological naturalism" by Anil K*. Seth, Behavioral.

Farisco, M., K., E., & Changeux, J.-P. (2024). Is artificial consciousness achievable? Lessons from the human brain. *Neural Netw*, *180*, 106714. https://doi.org/10.1016/j.neunet.2024.106714.

Godfrey-Smith, P. (2023). *Nervous Systems, Functionalism, and Artificial Minds*.

Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *NeuroReport*, *11*(11), 2351–2355.

Hills, P. J., & Lewis, M. B. (2006). Reducing the own-race bias in face recognition by shifting attention. *Quarterly Journal of Experimental Psychology*, *59*(6), 996–1002. https://doi.org/10.1080/17470210600654750.

Hobbes, T. (1651). *Leviathan*.

Huxley, T. (1894). *Evolution & Ethics*.

Khamassi, M., Nahon, M., & Chatila, R. (2024). Strong and weak alignment of large language models with human values. *Scientific Reports*, *14*, 19399. https://doi.org/10.1038/s41598-024-70031-3

M., F., P., S., & K, E. (2026). Functionalism in AI consciousness: ethical reflections on the relevance of the biological body. In M. Fay, F. Flother, & C. H. Hoffman (Eds.), *Computers with Salaries and Cemeteries: AI Ethics from Industry to Philosophy to Science Fiction*. Springer-Nature.

Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness*, *08*(01), 43–66.

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.

Michel, C., Rossion, B., Han, J., Chung, C. S., & Caldara, R. (2006). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, *17*(7), 608–615. https://doi.org/10.1111/j.1467-9280.2006.01752.x

Ortega, & Gasset. (1962). Man the Technician". In *History as a System and Other Essays toward a Philosophy of History*. W.W.Norton.

Phelps, E. A., Cannistraci, C. J., & Cunningham, W. A. (2003). Intact performance on an indirect measure of race bias following amygdala damage. *Neuropsychologia*, *41*(2), 203–208. https://doi.org/10.1016/S0028-3932(02)00150-1.

Ricœur, P. (1992). *Oneself as Another*. University of Chicago Press.

Seth, A. K. (2024). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, *48*, 267. https://doi.org/10.1017/S0140525X23001291.

Singer, P. (2025). *Isaac Asimov's Laws of Robotics are Wrong*. The Brookings Institution.

Spencer, H. (1851). *Social Statistics, 1851*.

Tennyson, A. (1850). In *Memoriam A.H.H*.

(2016). The War on Women.

Evers, K. (2026). Artificial Consciousness: Science Fiction, Utopia, or Pandora's Box? In *Proceedings of the Paris Institute for Advanced Study* (Vol. 21). https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box
2026/4 - paris-ias-ideas - Article No.1. Disponible https://paris.pias.science/article/artificial-consciousness-science-fiction-utopia-or-pandora-s-box - ISSN 2826-2832/© 2026 Evers K.