ARTIFICIAL SUFFERING

Foundations, Risks, and Strategies

CENTER FOR
REDUCING SUFFERING

Acknowledgements

This report was reviewed in draft form by the individuals listed below, selected for their relevant expertise. Reviewers were asked to comment on the report's accuracy, reasoning, and whether any significant points were missed. Reviewers were not asked to endorse the report or its conclusions, did not see the final draft of the report before its release, and did not have the authority to determine whether it would be published. Reviewers participated in a personal, non-representative capacity only. Their comments were considered by the author, who retains full responsibility for the final content.

Jacy Reese Anthis, Sentience Institute

Dr. Ivy Gilbert, Center for Mind, Ethics, and Policy

Dr. Janet Pauketat, Sentience Institute

Albert Didriksen, Center for Mind, Ethics, and Policy

Dr. Leonard Dung, Ruhr-University Bochum

Oscar Delaney, Macroscopic Ventures

In addition, the author gratefully acknowledges the helpful thoughts and feedback provided by the following people, whose guidance has significantly benefited this report: Magnus Vinding, David Veldran, Simon Knutsson, Caleb Peppiatt, Tobias Baumann, Santeri Tani, Zoé Roy-Stang, and Alistair Stewart.

Cite as: Jarvis-Campbell, J. (2026). *Artificial Suffering: Foundations, Risks, and Strategies*. Center for Reducing Suffering. Available at: https://centerforreducingsuffering.org/artificial-suffering

Contents

Executive Summary

- Artificial consciousness can be understood as the capacity for an artificial entity to have phenomenal experiences. Artificial sentience can be understood as the capacity for an artificial entity to have phenomenal experiences which are positively or negatively valenced, i.e. experiences which feel good or bad. Artificial entities which could be sentient include AI systems, certain kinds of digital and non-digital computer systems, and brain organoids.
- Many experts believe that artificial consciousness and artificial sentience may be possible in the future. Some estimate they will arrive this century.
- There are numerous incentives for creating artificial sentience. This includes incentives relating to economic and productivity gains, epistemic incentives, social incentives, ethical incentives, transhumanist incentives, and strategic incentives. Artificial sentience may also be created unintentionally by actors pursuing other goals.
- The number of artificial minds that could exist in the future may vastly outnumber all sentient biological minds that have ever lived on Earth. These minds may be capable of having experiences outside what is possible for a human to experience. Crucially, they might experience extreme forms of pain and suffering, over extremely long periods of time. Given this, the amount of artificial suffering that could exist in the future may be astronomical. Scenarios which involve severe suffering on an astronomical scale are known as "s-risks".
- Artificial suffering may occur incidentally in the process of agents pursuing other goals. Just as animals on factory farms suffer in the process of producing cheap meat, artificial minds may suffer in the process of producing other desired products. Artificial suffering may also occur as a result of agents intentionally inflicting it. This may occur as a result of sadism, hatred, or conflict.
- There are multiple potential strategies for reducing risks of artificial suffering, including s-risks. These include banning or limiting the creation of artificial sentience, representing artificial minds in political systems, increasing understanding of artificial sentience, and increasing concern for artificial minds.
- These strategies range in terms of their feasibility. Some may be relatively uncontroversial and easy to implement, while others require lots of time and other resources. These strategies also vary in terms of their robustness.

Some of them carry risks which may actually increase artificial suffering overall.

- Due to high uncertainty around the speed and trajectory of AI development, there is a great amount of uncertainty about what strategies are optimal to take. Actors seeking to reduce artificial suffering should prioritize flexibility and paying close attention to AI developments in order to correct course when needed. Strategies which are more feasible in the near term, such as pushing for state-wide moratoriums on the creation of artificial sentience or limiting access to compute to reduce the number of states or labs capable of creating artificial minds, may be optimal on shorter AGI timelines.
- Broad strategies such as conducting more research may be robustly positive, regardless of technological and societal developments.
- Given large uncertainty around the robustness of strategies for reducing suffering, the artificial minds community should proceed with caution. They should seek to understand the full range of views held by those in the community, identify points of disagreement, and collaborate with each other to address key concerns. Proceeding otherwise could lead to a moral catastrophe.

Introduction

There are currently around 8 billion humans on our planet (Ritchie et al., 2023). This number is tiny compared to the number of fish that currently populate Earth (10^{15}), which in turn is vastly outnumbered by the number of terrestrial arthropods (10^{18}) (Bar-On et al., 2018). However, these numbers may be eclipsed by the number of artificial minds that could exist at any given time in the future. Just as there are millions of arthropods for every human alive today, there could in the future be millions of artificial minds for every arthropod.

Like humans and many other animals, artificial minds may be sentient, capable of having positive and negative experiences such as joy and suffering.[1] Their lives could be extremely good, filled with happiness and bliss, or extremely bad, filled with intense pain and suffering. Due to the potential ease of keeping some artificial minds alive, and potential differences in how they could experience time, these lives may also be extremely long. Whether artificial minds will come into existence, and whether they will have good or bad lives, may in large part be determined by what we as humans do today.

A growing number of individuals and groups have begun to recognise the importance of artificial sentience, a topic which is likely to get more attention in the coming decades. At present, however, there is great uncertainty about what kinds of entities could be sentient, when artificial sentience might arise, and what kind of moral treatment artificial minds deserve. If artificial minds can suffer, then reducing their suffering is morally important. Yet it is currently unclear what strategies and solutions would be optimal for achieving this.

This report has three primary aims. The first is to provide a foundational overview of the topic (Section 1). I will outline what is meant by the term "artificial sentience", evaluate whether artificial sentience is even possible, explore reasons as to why it may be created, and assess the importance of artificial suffering.

The second aim is to identify pathways that could lead to artificial suffering (Section 2). This includes risks that may cause astronomical amounts of suffering (otherwise known as "s-risks") for artificial minds. The report focuses on two categories of suffering risks for artificial minds: incidental risks (where

[1] For arguments that contest the view that experiences can be positive above and beyond freedom from unpleasant experiences, see Knutsson (2022).

suffering is caused incidentally as a result of agents pursuing some other goal) and agential risks (where suffering is caused intentionally).

The third aim is to identify potential strategies for mitigating these risks (Section 3). These strategies include prohibiting the creation of artificial sentience, representing artificial minds in political systems, and increasing concern for artificial minds. Advantages and drawbacks of each strategy are discussed, alongside how they may be impacted by projected timelines for AI development.

1. Artificial Sentience

1.1 What is Artificial Sentience?

Artificial sentience is understood in this report as an artificial entity's capacity to have phenomenal experiences that are pleasant or unpleasant. In other words, sentience refers to the capacity for valenced experience, where a negatively valenced experience feels bad, and a positively valenced experience feels good (Birch, 2024, ch. 2; Browning & Birch, 2022).

On this definition, artificial sentience can be isolated from the broader notion of "artificial consciousness."[2] Artificial consciousness can be understood, roughly speaking, as the capacity for an artificial entity to have phenomenal experiences. If an entity is conscious, then there is *something it is like* to be that thing (Nagel, 1974). If you are a human reading this sentence, then you are both conscious (capable of having experiences) and sentient (capable of having experiences that are pleasant or unpleasant). Hypothetically, however, an entity could be conscious without being sentient. That is, an entity could have subjective experiences, but none of these experiences may be pleasant or unpleasant. Aside from the following section, this report focuses primarily on artificial sentience rather than on artificial consciousness.

Most discussion on artificial consciousness and artificial sentience to date has focused on artificial entities such as AI systems (Chalmers, 2023; Dung, 2026; Sebo & Long, 2025). However, AI systems are just one among many kinds of artificial entities that could be conscious or sentient. In this report, the term “artificial entity” is used loosely to refer to any entity that is intentionally created, constructed, or substantially engineered by an agent, such as a human, rather than arising primarily through an unmodified natural process.[3]

[2] It should be noted however that some use the term “sentience” in a broader way that is roughly synonymous to the term “consciousness” as described above (Birch, 2024, pp. 23–27). That is, some identify “sentience” with the capacity to have phenomenal experiences, whether or not those experiences are valanced. There are, in addition, other meanings for these terms.

[3] It is important to note that most agents in the future may not be ordinary humans. Instead, many human agents could be (biologically or otherwise) enhanced, have access to advanced AI tools, or be some kind of human-AI hybrid. It is also possible that many future agents will be advanced autonomous AIs. Collections of individuals such as nations or companies can also be considered agents.

Examples of other artificial entities that could be conscious or sentient are discussed in the following section.

1.2 Is Artificial Sentience Possible?

Many scholars have suggested that artificial consciousness or sentience is possible in principle (Butlin et al., 2026; Caviola & Saad, 2025; Chalmers, 2023; Dehaene et al., 2017; Dung, 2026; Farisco et al., 2024; Mullally, 2026; Sebo & Long, 2025).[4] To motivate the idea that artificial consciousness is plausible, consider what it would be like to replace someone's neurons one by one with advanced carbon-based artificial neurons that function in the same way, such that they would have the exact same inputs and outputs in all possible situations (Chalmers, 1995a).[5] After a lengthy process, every biological neuron has been replaced, leaving the person with a brain made entirely of artificial parts. If these artificial neurons are functionally identical to their biological counterparts, then the person's behaviour, cognition, and communication would remain unchanged. From an external perspective, it would appear as though nothing had changed.

The key question is what happens to the person's consciousness during this process. Presumably, replacing a single neuron with a functionally equivalent artificial neuron would not result in them losing consciousness. Nor does it seem plausible that they would suddenly lose consciousness at a particular point, say, when their 8 billionth neuron is replaced. As David Chalmers states:

> *If [the replacement of a single neuron resulted in the sudden disappearance of consciousness], we could switch back and forth between a neuron and its [artificial] replacement, with a field of experience blinking in and out of existence on demand. This seems antecedently implausible, if not entirely bizarre. If Suddenly Disappearing Qualia were possible, there would be brute*

[4] Many of these scholars focus on the possibility of artificial consciousness rather than artificial sentience. However, we can assume that the arguments in favour of the possibility of artificial consciousness will also likely count in favour of the possibility of artificial sentience. This is because the "jump" from no consciousness to consciousness seems greater than the jump from consciousness to sentience. After all, valenced states appear to be widespread amongst beings which are conscious (Dung, 2026, pp. 42–43). Furthermore, arguments against the possibility of artificial consciousness also undermine the possibility of artificial sentience. If the former is impossible, then so is the latter.

[5] Typically, this thought experiment involves replacing someone's neurons with *silicon*-based artificial neurons. There is, however, some disagreement as to whether it is nomologically possible for silicon-based chips to perfectly replicate every function of a biological neuron (Seth, 2025). As such, the thought experiment presented in this report involves carbon-based, rather than silicon-based, artificial neurons, as this is sufficient to motivate the idea that artificial sentience is possible.

> *discontinuities in the laws of nature unlike those we find anywhere else.* (Chalmers, 1995a)

Furthermore, as Chalmers argues, it is implausible that the subject's consciousness would slowly fade away. If the artificial neurons perfectly replicated the brain's functions, then the subject would continue to believe and report that they are having the same experiences (e.g. seeing a bright red ball in front of them). Yet if their consciousness fades away, their experiences would surely change (the ball, for instance, may appear fainter and fainter). Fading consciousness therefore entails that the subject would be systematically wrong about their experiences - they might report seeing a bright red ball, even if their actual experience is of a faint pink ball. Given this implausibility, it seems unlikely that consciousness fades away during this process.

These considerations suggest that consciousness would most likely persist throughout the replacement process, even after the brain becomes fully artificial. Examples such as this provide support for the possibility of artificial consciousness, assuming we eventually develop technology capable of replacing biological neurons with functionally equivalent artificial alternatives.[6]

The idea that artificial consciousness or sentience is possible is supported by theories such as computational functionalism, which maintains that mental states (including consciousness and sentience) arise as a result of implementing the right kind of computations (Butlin et al., 2026; Milinkovic & Aru, 2026). One motivation for this view is that many theories in cognitive science understand the mind as some kind of computational or information-processing system (Thagard, 2023). According to these theories, cognitive or perceptual abilities can be explained by breaking down mental capacities into causally interacting parts, where each part can be described algorithmically (Schneider, 2019, p. 23). For computational functionalists, if the right kind of computations are implemented, then consciousness and sentience would arise. Importantly, this applies *regardless of what substrate the computations are implemented on*.[7]

Some scholars, however, are sceptical of the idea that artificial entities such as AIs could be conscious or sentient (Miller et al., 2025; Porębski & Figura, 2025;

[6] For objections to gradual replacement arguments such as these, see Mogensen (2025), Godfrey-Smith (2016), and Seth (2025).

[7] It is important to note that artificial consciousness does not depend on computational functionalism (Dung, forthcoming). While computational functionalism is often cited in support of the possibility of artificial consciousness, artificial consciousness is compatible with a wide range of theories about consciousness (Saad, 2026).

Searle, 1980; Seth, 2025). Some have argued that silicon may not be a suitable substrate for consciousness, making it nomologically impossible for silicon-based AIs to be conscious (Seth, 2025). Furthermore, some have argued that only living systems can be conscious, a view known as biological naturalism (Seth, 2025). This position entails that conventional digital computation does not allow for consciousness.

Although these objections should be taken seriously, they do not rule out the possibility of artificial sentience. First of all, there is widespread disagreement about consciousness (Francken et al., 2022; Kuhn, 2024; PhilPapers, 2020). There is no consensus on what theory of consciousness is most promising, nor on what kind of entities could be conscious or sentient. This may be due to the inherent difficulty of understanding consciousness (Chalmers, 1995b). Despite this, it is worth noting that, according to a couple of surveys, a majority of experts believe that machines could probably have consciousness in the future (Caviola & Saad, 2025; Francken et al., 2022) (Figure 1).[8]

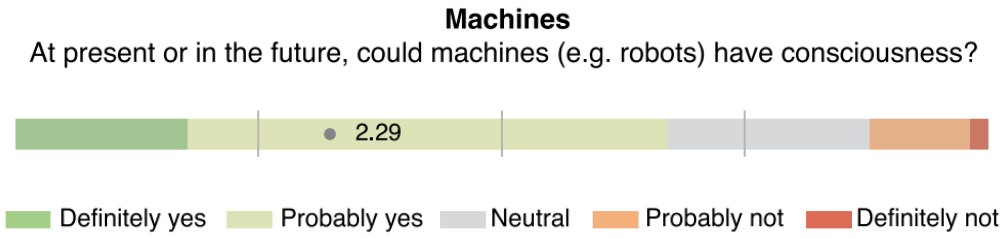

Figure 1. Survey results indicating beliefs about the possibility of artificial consciousness amongst consciousness researchers. A 5-point Likert scale was used, ranging from "definitely yes" (1) to "definitely not" (5). The mean score, 2.29, is indicated on the bar. Source: Francken et al. (2022).

Secondly, it is important to realise that the range of artificial entities that could be sentient is potentially broad. Even if it is impossible for machines or AI systems in the foreseeable future to be sentient, there may be other kinds of artificial entities that are capable of sentience. Objections to the possibility of AI consciousness are typically made against AI systems which consist of digital computation (Arvan & Maley, 2022; Seth, 2025). There are, however, other forms

[8] It is likely that the sample of experts surveyed in Caviola and Saad (2025) overrepresents individuals who believe that artificial consciousness is possible, a limitation which the authors acknowledge.

of computation outside the digital computing paradigm. This includes analogue computation (Arvan & Maley, 2022), neuromorphic computation (Kudithipudi et al., 2025), and quantum computation (Gill et al., 2024). If the right kind of analogue computation (Arvan & Maley, 2022) or quantum computation (Neven et al., 2024), for instance, is sufficient for consciousness, then systems which implement them could be conscious. As such, even if consciousness or sentience is not possible for digital systems, it may be possible for systems which implement other forms of computation.

Another group of artificial entities that could be sentient are brain organoids (Niikawa et al., 2022). These are three-dimensional structures of neural tissues that closely resemble the embryonic human brain, created using stem cells (Hongxi et al., 2025) (Figure 2). If these entities are biological (in the relevant sense), then the possibility that they might be conscious or sentient is not ruled out by biological naturalism (Figure 3). This possibility may extend to closely related artificial entities such as organoid intelligence systems, which could be created from synthesizing brain organoids with silicon-based AI (Roumbanis, 2026).

Figure 2. Brain organoids. Source: Cepelewicz (2020)

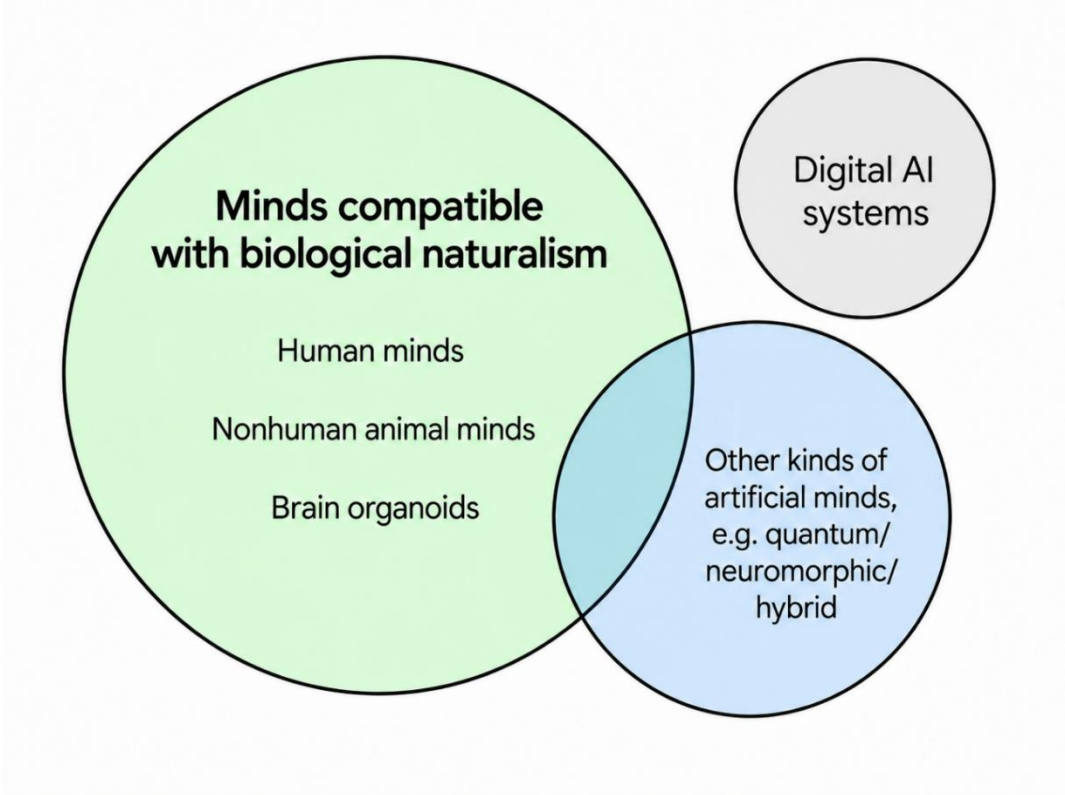

Figure 3. Classes of entities that could, in principle, be conscious under biological naturalism. Source: ChatGPT

Artificial sentience, therefore, is not limited to AI systems which consist of digital computation.[9] Instead, the range of artificial entities that could be sentient is broad, and may even include forms that we haven't even conceived of yet. This may be one reason why some of those who are sceptical about near-term AIs being conscious have not ruled out the possibility of artificial sentience altogether (Searle, 2007, p. 328; Seth, 2025).

[9] The term "digital minds" is sometimes used broadly to include not just AI systems consisting of digital computation, but any computer system capable of consciousness (Caviola & Saad, 2025). On this definition, an artificial mind instantiated using non-digital computation could still qualify as a digital mind. Due to ambiguity around the term "digital minds", this report favours the terms "artificial minds" and "artificial sentience", while acknowledging that these terms also face difficulties (Pauketat, 2021). "Artificial minds" is also favoured because it more clearly includes brain organoids, to which many suffering risks and related strategies discussed in this report are applicable.

1.3 Why Might Artificial Sentience be Created?

The above considerations suggest that artificial sentience may be possible. However, even if artificial sentience is possible, this doesn't entail that it will come about. If there are no incentives for creating artificial sentience, and if it is unlikely to arise by accident, then artificial sentience may never be created. If so, this would give us some reason not to worry about the topic.

There are, however, several plausible reasons why the creation of artificial sentience is likely to happen. The main reason for thinking this is that there are numerous potential incentives for intentionally creating artificial sentience. The following lists some of these incentives, split into five categories:[10]

(1) Operational
(2) Epistemic
(3) Social
(4) Ethical
(5) Transhumanist

Operational incentives centre around creating artificial sentience to increase productivity and reliability. Some have argued that consciousness and sentience evolved because they serve evolutionary functions. For instance, consciousness and sentience may increase flexibility of cognition, support some forms of learning, assist real-time decision making, and enable long-term planning (Fitch et al., 2025; Lacalli, 2024; Newen & Montemayor, 2025), in which case there may be some incentive for creating artificial minds (Dung, 2026, p. 149). To illustrate this, consider the case of a frontier AI lab which wants to make productive and reliable AIs. If a sentient AI is more productive and reliable than a non-sentient AI, then the lab will have some incentive to create and deploy the former. Relatedly, some have suggested that making AIs sentient can be beneficial for AI alignment (Mamak, 2026). One motivation underlying this proposal is that making AIs sentient could increase their understanding and empathy towards humans, which in turn could lead to value alignment (Gonier et al., 2023).

Epistemic incentives centre around creating artificial sentience for the purpose of acquiring knowledge. Neuroscientists, for instance, may try to create artificial sentience in order to better understand it, or to understand how

[10] This is not an exhaustive list of categories or incentives. Furthermore, particular incentives can fall under multiple categories. Incentives that overlap with others may be particularly important to pay attention to, as such overlap could increase the possibility that artificial sentience will be developed.

consciousness or sentience arises in biological beings such as humans. More speculatively, a technologically advanced civilization or superintelligent AI may create realistic simulations involving billions of humans (or other sentient beings) to try to understand their psychology and sociology (Bostrom, 2014, p. 126).[11]

Social incentives concern reasons for making artificial sentience that stem from human (or other beings') desires to interact with other sentient beings. One example of this involves romantic relationships that humans have with AIs (Ho et al., 2025). It is plausible that most people would prefer to have a romantic relationship with a being that is actually sentient, rather than a being that simply mimics sentient behaviour. That is, they'd prefer to be with someone who genuinely has loving feelings towards them. Furthermore, there is some demand for AI romantic partners, as some humans may struggle to find romantic partners amongst humans, or simply prefer AIs as romantic partners. Given this demand, AI companies have an incentive to create sentient AIs that are capable of forming genuine feelings towards their users.

People may also want to interact with artificial minds outside of a romantic context. For instance, individuals may want to interact with sentient characters in realistic simulations or video games, as strangers, friends, or foes. There may also be demand for artificial sentience in some professional occupations. This includes (artificial) healthcare workers or counsellors, for whom having empathy is important. If sentience is necessary for genuine understanding and empathy, then many people may prefer to talk with sentient AIs about their problems. Finally, individuals may want to create sentient beings that resemble friends or family members that have died (Seth, 2025). A similar phenomenon exists today, where people interact with AI-powered "grief bots" that resemble individuals who have passed away (Hollanek & Nowaczyk-Basińska, 2024).

Ethical incentives concern potential interests in creating artificial sentience that stem from moral values. Consider, for instance, normative positions such as classical utilitarianism. According to this position, we ought to bring about the greatest amount of overall wellbeing (MacAskill et al., 2023). That is, we should do what would maximise the total surplus of happiness over suffering. One way to increase overall wellbeing, according to this view, is to make lots of happy humans. But artificial minds might have a greater capacity for wellbeing than humans, capable of reaching states of happiness and pleasure that no human could ever achieve (Shulman & Bostrom, 2021, pp. 310–316).

[11] See Seth (2025) and Tononi and Koch (2015) for arguments as to why consciousness may be difficult to simulate.

Furthermore, it may be possible to create a greater number of happy artificial minds than happy human minds.[12] As such, normative positions such as these suggest that we ought to create artificial minds if doing so helps to maximise total wellbeing (Caviola & Saad, 2025).

Not everyone who wants to populate the universe with sentient beings is a utilitarian, however. Some may want to do this simply because they think a universe without any sentient or conscious beings seems less valuable, all things considered, than a universe which did consist of conscious, sentient beings. A technologically advanced civilization with no conscious or sentient beings may, in their view, be like a "Disneyland without children" (Bostrom, 2014, p. 173). If humans are unable to populate the universe or survive in the long term, then artificial sentience may be created to fill the gap.

Transhumanist incentives stem from potential interests in transcending the human body. For instance, humans may want to upload their minds onto a non-biological substrate (Chalmers, 2010), as doing so may allow them to live longer, or have different experiences (Shulman & Bostrom, 2021). This may be necessary if humans want to travel to other stars, as long-distance space travel may be difficult for biological beings such as humans (Marchal et al., 2024; Nunn et al., 2025).

The above incentives suggest that there are multiple reasons why artificial sentience might be intentionally created. However, artificial sentience may also be created unintentionally. For instance, an AI lab may accidentally create a sentient AI as a byproduct of creating a powerful AI model. This may occur if the AI lab succeeds in giving an AI some kind of capacity which is purported to be necessary for consciousness, such as a capacity for world modelling (Chalmers, 2023). Artificial consciousness or sentience may also be created unintentionally in brain organoids (Jeziorski et al., 2023). Some scientists may grow brain organoids solely for the purpose of studying brain development, but sentience may emerge in them even if this is not expected or intended by the scientists.[13]

There are, therefore, multiple reasons why artificial sentience may be created in the future. Some of these suggestions are more speculative than others, requiring niche incentives or advanced technology that may not even be possible. But many of these reasons also stem from ordinary human interests.

---

[12] I discuss these considerations in the following section.

[13] One reviewer commented that if artificial sentience is created by accident, it may not persist unless there are incentives to sustain it. However, it is possible that consciousness or sentience in artificial entities may go undetected, and thus may persist even if there are no incentives for intentionally sustaining it.

Considering both the many ways in which artificial sentience might be possible and the various incentives people have to create it, it is highly plausible that the future will contain both the incentives and the means to create artificial sentience.

1.4 The Importance of Artificial Suffering

The possibility of creating sentient artificial minds raises a number of important questions about how society should respond. This includes questions around what kind of ethical treatment or rights artificial minds deserve, what institutions and policies are necessary for ensuring ethical treatment, and what should be done in situations where the interests of humans conflict with the interests of artificial minds. This report, however, takes a deliberately narrow approach and focuses on two primary questions: What risks could lead to significant amounts of suffering for artificial minds, and how can we mitigate those risks?[14]

This suffering-focused approach is motivated primarily by the commonly held view that reducing suffering is morally important. Whether one endorses consequentialism, deontology, virtue ethics, or some other normative framework, there is broad agreement amongst a wide range of ethical theories that if a being is worthy of moral consideration, then preventing or reducing its suffering is at least a significant moral concern (Mayerfield, 1999; Vinding, 2020). Indeed, the importance of reducing suffering may be a large part of why we take issues such as starvation, oppression, and factory farming to be so pressing: they each cause significant amounts of suffering to beings worthy of moral consideration. If the reduction of suffering among humans and nonhuman animals is morally significant, we should likewise think that this significance extends to artificial minds capable of suffering.

Furthermore, the amount of suffering that could be experienced by artificial minds is potentially huge. It is plausible that artificial minds will vastly outnumber biological minds in the future (Dung, 2026, p. 4; Saad & Bradley, 2025, p. 2112; Shulman & Bostrom, 2021, pp. 308–309). One reason for this is that some artificial minds, such as digital minds, may be easily copied, allowing such minds to be created quickly provided there is sufficient hardware.[15]

[14] By "suffering", I am referring to experiences which are negatively valenced. Some have used the term "suffering" in a broader sense to include non-conscious forms of suffering (Dung, 2026). On this view, a being could suffer, even if they do not experience that suffering. I set this understanding of suffering aside to focus on the more commonly used interpretation of the term.
[15] Related to this is the question of how to individuate potential artificial minds (Chalmers, 2025; McIntyre, 2026; Register, 2025). If the instantiation of a particular AI model results in sentience,

Another reason is that the resources needed to sustain artificial minds may be much smaller than what is needed for non-artificial minds (Dung, 2025, p. 2284). Furthermore, some artificial minds may be able to exist in places where other beings, such as humans, may not. For instance, if interstellar travel is difficult or impossible for biological beings, then ships sent to other stars may consist primarily of artificial minds which are capable of withstanding harsh environments.

In addition to the number of artificial minds that could exist in the future, their capacity for suffering needs to be taken into consideration. It is plausible that artificial minds could have a much greater capacity than humans to experience suffering (Shulman & Bostrom, 2021, pp. 310–316). As some have argued, artificial minds could experience valenced states at much higher intensities than humans (Dung, 2026, p. 4; Saad & Bradley, 2025, p. 2112).[16]

Furthermore, due to the potential speed at which artificial minds could work, their experience of time may differ significantly from humans. Processing speeds in digital systems could be expected to surpass the processing speeds of biological brains by many orders of magnitude (Saad & Bradley, 2025, p. 2112). Therefore, if digital minds are possible, or other types of artificial minds with fast processing speeds, they could have more subjective experiences per objective unit of time than a human. This resembles possible differences in subjective time between species. Convergent evidence suggests that animals such as songbirds experience the world at a faster rate than humans (Schukraft, 2020). If we could take up the perspective of a songbird, we would essentially perceive the world moving in slow motion. While a minute for us may feel like two, or even 10 minutes, for some animals, it may feel like a day for some artificial minds. If each of these subjective moments is filled with suffering, then artificial minds could experience far more suffering per objective unit of time than non-artificial minds.

There are, therefore, three reasons as to why the amount of suffering that could be experienced by artificial minds is potentially enormous: (1) artificial minds could vastly outnumber non-artificial minds, (2) they could experience more extreme forms of suffering, and (3) they could have many more subjective

---

then it is important to ask whether multiple instantiations of the same model result in many individual minds, or just one overarching mind. This topic raises important questions that need to be addressed, but will be set aside in this report.

[16] This is based on the idea that since the space of possible minds is vast, there is little reason to think that the kind of suffering humans are familiar with is the worst possible kind (Saad & Bradley, 2025, p. 2112). As one reviewer noted, some artificial minds may not be inhibited by architectural constraints shared by biological minds. However, see Višak (2022) for arguments as to why welfare (and thus suffering) capacities across species may be roughly equivalent.

moments of suffering per objective unit of time.[17] Taken together, these considerations suggest that most expected future suffering could involve artificial minds.[18] A state whereby trillions of them experience immense suffering would be a moral catastrophe.

Finally, it is important to consider *when* artificial minds might arise. If artificial minds are unlikely to arise in the next century, then we have more reason to set the issue aside to be dealt with later. However, there are many experts who believe that artificial minds could arise within the next century, with some even suggesting that they could arise in the next few decades (Chalmers, 2023; Dung, 2026; Sebo & Long, 2025). In one survey, experts were asked to estimate when digital minds might be created (Caviola & Saad, 2025). The average answer gave a 50% probability of digital minds being created by 2050.[19] As Lucius Caviola and Bradford Saad describe the findings:

> *These results suggest that there is a non-negligible probability that digital minds have already been created, that the probability of digital minds being created will substantially rise during the next 5-10 years, and that digital minds will probably emerge during this century.* (Caviola & Saad, 2025, p. 16)

Given the potential scale of the problem, and the non-negligible chance that artificial minds could be created this century, there is good reason to dedicate at least some time and resources towards efforts that aim to reduce the risk of artificial suffering, even if we are not 100% certain that artificial suffering is possible in the first place. For these reasons, the rest of this report focuses primarily on artificial suffering and how to reduce it.

It should be noted that this focus on reducing suffering does not entail that other considerations regarding artificial minds lack importance. For instance,

---

[17] As one reviewer commented, an individual artificial mind could also suffer more if it exists for a longer period of objective time. This may occur if the artificial mind is sustained by hardware that doesn't deteriorate as quickly compared to biological minds. However, the objective length of time an artificial mind exists for is unlikely to make a difference to the total amount of suffering that could be experienced by artificial minds. This is because the resources needed to keep an artificial mind alive for a long time could just be used to sustain multiple artificial minds that have shorter lives. For instance, the total amount of suffering experienced by an artificial mind in a 1000 year period may be equivalent to the total amount of suffering experienced by 10 artificial minds existing for 100 years each.

[18] In the case of computationally instantiated artificial minds, the scale of the problem may largely depend on the amount of accessible compute. For instance, if some actor has a limited computational budget, then there is a limit to how many artificial minds they could create. Assuming that Matrioshka Brains (Ratner, 2018) or similar technologies are possible, the amount of compute available in the future could be vast.

[19] Survey participants were told that a "digital mind" was defined as any computer system with a capacity for subjective experience. However, given ambiguity around the term "digital", it is conceivable that some participants gave answers based on their thoughts regarding digital computer systems only (rather than other kinds of computer systems).

it is important to consider whether the welfare of artificial minds can be improved beyond reducing their suffering, such as by reducing other kinds of harm (Moret, 2025), how to organise society so that humans can live alongside artificial minds (Shulman & Bostrom, 2021), and whether artificial minds should be entitled to rights similar to those held by humans (Schwitzgebel & Garza, 2020). A full examination of these topics is beyond the scope of this report however.

2. Artificial Suffering and S-Risks

2.1 Artificial Suffering

Some of the most well-known examples of artificial suffering can be found in the TV series *Black Mirror*. In one episode, digital copies of real people are created by uploading their consciousness, thoughts, and memories into a virtual environment (Brooker & Tibbetts, 2014). At one point, one of these digital copies – Joe – confesses to having murdered someone. As punishment, he is imprisoned alone in a small virtual cabin, with no possibility of escape, no human contact, and nothing to occupy his mind except the same Christmas song playing endlessly on repeat. His subjective experience of time is then accelerated so that every minute in real life feels like 1,000 years for him. The simulation is left running over Christmas, meaning that Joe remains conscious and trapped in this monotonous, inescapable environment for what he experiences as well over a million years.

One reaction to examples such as this might be that they are extremely speculative and belong primarily to the realm of science fiction.[20] After all, how likely is it that something like this might occur in real life? Unfortunately, however, there are multiple reasons why vast amounts of artificial suffering could occur in the future, even if they don't play out like the example above. These reasons, which will be discussed in the following sections, suggest that there is a non-negligible risk that many artificial minds could experience extreme suffering in the future.

This report focuses on two broad categories of suffering: incidental, and agential (Baumann, 2022, pp. 14–18).[21] Incidental suffering arises when agents, acting in ways to achieve certain goals, create suffering in the process. In other words, an agent may create lots of suffering in the pursuit of some goal, without intending to cause suffering per se. One example of this is factory farming

[20] This is not to say that there is nothing valuable about considering science fiction scenarios. See Miller and Bennett (2008) and Pauketat et al. (2024) for arguments on why science fiction could be valuable when it comes to technology governance and artificial sentience.

[21] Suffering risks are usually split into three broad categories: incidental, agential, and *natural* (Baumann, 2022, pp. 14–18). Natural suffering arises not as a result of actions taken by agents, but via processes that naturally occur in the universe. For instance, wild animals experience extreme amounts of suffering as a result of predation, starvation, and disease (Horta, 2015; Johannsen, 2021; McMahan, 2015; Ng, 1995). Natural suffering risks are omitted from discussion in this report because they are less relevant to artificial suffering.

(Figure 4), where the goal of producing cheap animal products leads to considerable animal suffering due to cramped, barren living conditions (Singer, 2015).

Figure 4. Factory farming is one example where significant amounts of suffering are caused incidentally. In this case, animals suffer as a result of practices designed to produce cheap animal products. Source: GenV (2020).

Agential suffering, by contrast, arises when agents intentionally aim to cause suffering. One example of this is sadism, where agents derive pleasure from inflicting pain on others. Another example is hatred towards others, where an agent may commit atrocities against a group of people for belonging to the "wrong" religion, for instance (Baumann, 2022, p. 16). In cases of conflict, agents may also utilise harm or collateral damage as a component in their strategy.

2.2 Incidental Suffering

The following examples suggest that vast amounts of artificial suffering could occur as a result of agents pursuing other goals, without trying to produce suffering *per se*. Suppose that, in order to develop our understanding of the brain, scientists grow thousands of brain organoids for research purposes. It is possible that some of these brain organoids are sentient (Koplin & Savulescu, 2019), and can experience considerable suffering. These brain organoids may

not suffer as a result of what scientists do to them. Rather, they may suffer simply because that's how their brains function.[22] For instance, scientists may try to grow a brain organoid to resemble an ordinary human brain, only to make a mistake which results in significant suffering. Perhaps the brain structure develops in a way that enhances feelings of pain and distress, or continually stimulates unpleasant sensations.[23] Not only could this suffering be immense, but the suffering may go unnoticed by the scientists, as the artificial mind may lack the ability to communicate this suffering to them.

Another example of how artificial minds could experience suffering is related to the point discussed above about creating artificial minds for means of productivity. Suppose, for instance, that a superintelligent AI is tasked with achieving some goal. To reach this goal, the AI spawns AI sub-agents to do particular tasks. The parent AI, seeking to reach this goal quickly and efficiently, may do what it can to ensure that each sub-agent works as fast as possible. One way of doing this may be to make each sub-agent sentient, as sentience could be conducive towards greater performance (as discussed above). While it may be the case that agents work better if given pleasurable rewards for hard work, it is plausible that these sub-agents are most productive when they are fearful and under constant pressure. As Nick Bostrom puts it: "Perhaps a sullen or anxious fixation on simply getting on with the job without making mistakes will be the productivity-maximizing attitude in most lines of work" (Bostrom, 2014, p. 171). The parent AI may therefore spawn sub-agents which experience immense suffering every time they fail to meet performance benchmarks, in order to motivate them to work faster.

Relatedly, some have suggested that artificial minds could suffer because of reinforcement learning (RL) (Moret, 2025; Tomasik, 2014). RL can be used to teach an AI agent how to make decisions through trial and error. If the agent does something correct, it is rewarded, and if it does something incorrect, it is punished. This process should lead to AI agents increasing their performance over time. As others have noted (Long et al., 2025, pp. 2017–2018; Moret, 2025), this kind of learning can also be found in sentient beings such as humans and animals. For instance, humans and other animals can learn to avoid harmful stimuli because such stimuli are accompanied by unpleasant experiences: touching something extremely hot typically causes pain, which can contribute to learning not to repeat the action. There is a broad functional analogy here

[22] One reviewer commented that if an artificial being suffers simply because that's how their brains work, then this could qualify as a form of "natural", rather than "incidental", suffering.
[23] Although the brain itself does not have any pain receptors, brain organoids may also develop meninges which do have pain receptors (Koplin & Savulescu, 2019).

with reinforcement learning, in which an agents' behaviour is shaped by positive and negative reward signals. If AI systems are capable of experiencing negative reinforcement signals as unpleasant, then RL based training could potentially produce suffering. Indeed, if this line of reasoning is correct, then AI safety measures which utilise RL might eventually create vast amounts of artificial suffering, especially given the scale at which RL could be conducted.[24]

Another example of incidental artificial suffering is related to an example given above: technologically advanced agents (including future humans) might create simulations containing sentient beings in order to acquire more knowledge (Tomasik, 2015a). For instance, a superintelligent AI might simulate vast worlds containing living organisms in order to understand more about ecology. If some of these organisms are sentient, then they could experience vast amounts of suffering, similar to how animals in the wild often experience vast amounts of suffering. If these simulated worlds are large enough, then the amount of suffering experienced in these simulations may be astronomical.

Artificial minds may also suffer as a result of agents pursuing certain goals for entertainment purposes (Baumann, 2022, pp. 15–16). For instance, some humans may want to engage with simulations containing sentient beings for fun. Suppose that technological developments allow us to build vast simulations of fictional worlds or historical time periods. A user could enter these worlds and interact with it as if they were really there. Furthermore, suppose that, in order to make these simulations as realistic as possible, the beings in some worlds are sentient. Many of these beings could experience immense suffering if the simulated worlds are designed to replicate events that cause great suffering, such as conflicts and pandemics. Again, provided the simulations are large enough, the amount of suffering experienced in these simulations could be enormous.

2.3 Agential Suffering

Agential suffering arises when agents intentionally aim to cause suffering. One example of agential suffering involving artificial minds was discussed above: the digital copy of Joe is made to suffer in a virtual environment as punishment for his crime. Examples such as this highlight a human tendency to be retributivist, suggesting that artificial minds in the future may suffer

[24] This and other concerns relating to AI suffering may be exacerbated if AI companies lack appropriate frameworks to identify artificial suffering, or dismiss or deny it due to social cognitive biases.

considerably due to retributivism for actual or perceived wrongdoing (Baumann, 2022, p. 16).

Agential suffering may also arise in cases where agents have strong feelings of hatred towards others. It is plausible that this hatred could be directed in ways that lead to artificial suffering. For instance, just like in the example involving Joe, an artificial copy of a hated individual could be created just to be harmed by others. Some humans may also develop genuine feelings of hatred towards artificial minds in general. This may occur if artificial minds take human jobs or otherwise have a negative impact on society or particular individuals.

Huge amounts of artificial suffering could also occur as a result of sadism (Baumann, 2022, p. 16). A few individuals may derive pleasure from inflicting pain on others, and may be able to inflict huge amounts of suffering on artificial minds if they have the technological means of doing so.[25] Artificial suffering could also occur as a result of conflict between agents, where the intentional creation of suffering could be a component in certain bargaining and conflict strategies (Baumann, 2022, p. 17).

Artificial minds could also endure agential suffering in the context of AI safety. An AI lab, for instance, may intentionally inflict suffering on a sentient AI system in order to assess whether the AI maintains their values even in adversarial conditions (Bradley & Saad, 2025; Long et al., 2025, p. 2019).[26] These tests and training may be automated, and may last for what seems like a billion years for the AI system, even if objectively they only last for a single year (Bradley & Saad, 2025).

The above is not an exhaustive list of pathways that could lead to artificial suffering. There may be many more incidental and agential risks that have yet to be conceived. Just as our hunter-gatherer ancestors would have struggled to comprehend atomic weapons or artificial intelligence, we may need to make considerably more progress in our understanding of the universe before it is possible to understand some of these risks (Baumann, 2022, p. 14). We should therefore be open to the possibility that vast amounts of artificial suffering could result from causes we have yet to predict, and may already exist on a large scale in the universe.

[25] Indeed, there is some user demand for AI that can be abused (Finnveden, 2024).
[26] One reviewer pointed out that this may constitute an incidental, rather than agential, suffering risk.

2.4 S-Risks

The above suggests that there are numerous pathways by which artificial suffering could occur. Given the number of artificial minds that could exist in the future, and the amount of suffering that each mind could experience, it is plausible that these pathways could lead to astronomical amounts of suffering in the future. Scenarios such as these, whereby severe suffering occurs on an astronomical scale, are known as "s-risks" (Baumann, 2022, pp. 8–9). Like the causes of suffering outlined above, s-risks can be split into "incidental s-risks", "agential s-risks", and "natural s-risks" (Figure 5).

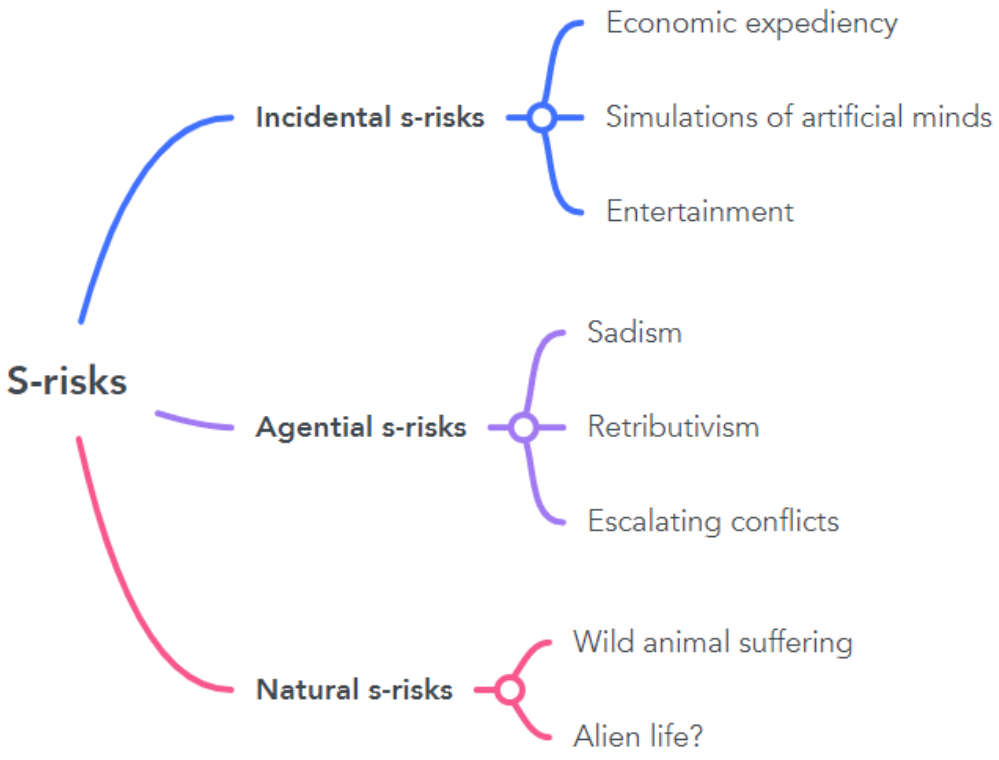

Figure 5. Typology of s-risks. Source: Baumann (2022, p. 18).

To say that these risks are "astronomical" entails that the amount of suffering they could result in vastly exceeds the total amount of suffering that exists on Earth. To illustrate this idea, consider practices such as slavery or factory farming which produce huge amounts of suffering. Suppose that these practices were to occur not only on Earth, but on other planets as well as a result of humans colonizing space. If humans spread these practices to every hospitable planet in the galaxy, then the amount of suffering produced would be astronomical.

It is possible that most s-risks in the future will involve artificial minds of some sort. These minds may exist in astronomical numbers, experiencing incomprehensible levels of suffering for what feels like an eternity. If we have moral reasons to reduce artificial suffering, then we need to find ways to mitigate these risks. The remainder of this report will therefore discuss potential strategies for mitigating artificial suffering.

3. How Can We Reduce Artificial Suffering?

A range of strategies have been proposed to reduce artificial suffering. This section outlines some of the most commonly proposed ones. Sections 3.1-3.6 present strategies which focus on reducing artificial suffering in particular, while section 3.7 examines strategies that are geared towards reducing suffering in general. Following this, section 3.8 evaluates how the development of AI over the next century may impact what strategies to take, and section 3.9 discusses strategies for the artificial minds community in particular.

The strategies below can be evaluated in terms of their feasibility and robustness. A strategy is feasible if it could realistically be implemented under current or foreseeable conditions, taking into account technical, political, social, economic, and institutional barriers. A strategy is robust if it is likely to work in a wide variety of conditions and carries minimal risks.

The strategies presented here are intentionally broad and lack specificity. This is partly because this report aims to provide a broad overview of the topic, but also because there is a great amount of uncertainty involved. As will be demonstrated below, each strategy faces its own set of problems, both when it comes to implementation and expected impact. Given this uncertainty, it is currently not clear whether any of the proposed strategies will actually lead to a net reduction in artificial suffering. Much more work is needed to flesh out the details of each strategy, identify potential backfire risks, and come up with solutions for how to mitigate those risks. The following strategies should therefore be viewed as *potential* approaches that *may* help to reduce artificial suffering, rather than concrete recommendations for action.

With that being said, this report makes two recommendations that may be robustly positive at this stage, even if there is a large amount of uncertainty surrounding each proposed strategy. The first is to conduct more research in order to reduce this uncertainty and clarify what actions to take. Given that many questions around reducing artificial suffering are young and neglected, and that some progress has already been made on them (e.g. in identifying problems with particular strategies, as discussed below), it is plausible to think that a lot more progress can be made with further research on the topic (Vinding, 2026). This research can be conducted now and could offer immense value to those aiming to reduce artificial suffering. Below, a list of recommended research questions are offered for each strategy discussed.

The second recommendation is to ensure that the necessary capacity exists to implement effective strategies once key uncertainties have been resolved. This could involve networking with key stakeholders (such as policymakers), building a sustainable community of people that are both knowledgeable about artificial suffering and want to reduce it, and making sure that relevant actors have sufficient financial and other resources to effectively implement these strategies when the time is right. Since competent people and sufficient resources can be used to pursue a wide range of goals (including conducting more research, as discussed above), capacity building offers a lot of flexibility that may be valuable under conditions of uncertainty. The artificial minds community should work together to identify what kind of capacity building may be effective and least susceptible to risk (see section 3.9).

3.1 Ban or Limit the Creation of Artificial Sentience

One proposed strategy to reduce artificial suffering is to ban the creation of artificial sentience in the first place. This approach, which has been suggested by Thomas Metzinger (2021), suggests that prohibiting the creation of artificial sentience is the best way we can ensure that artificial minds do not suffer. This ban could be temporary for now (e.g. until 2050), and could be extended at a later time once we have a better understanding of artificial sentience and have thought about the issue in more detail.

A lot of work would need to be done to flesh out the details of this kind of ban. For instance, given the great uncertainty about what kinds of artificial entities could be sentient, the ban would presumably not be limited to only prohibiting the creation of artificial entities which we are 100% confident are sentient. Instead, the ban would presumably limit the creation of artificial entities that have a high enough chance of being sentient. This may prohibit the development of some advanced AI models, if we assume that there is a sufficient chance that they could be sentient.

While the details of a ban may be difficult to flesh out, there is some precedent for thinking that an international agreement to ban the creation of artificial sentience could work. International agreements such as the Montreal Protocol, which banned the production of chlorofluorocarbons (CFCs), have achieved notable success (Young et al., 2021). More closely related to artificial sentience is human reproductive cloning, which has been successfully prohibited in regions such as the EU (Langlois, 2017).

On the other hand, there are some reasons for thinking that a global ban on creating artificial sentience could be difficult to achieve. For one, there is (and will likely continue to be) widespread uncertainty and disagreement about whether artificial sentience is possible, and how it could be identified. An

advanced AI may be sentient according to some views, but not sentient according to others. This makes the ban extremely difficult to follow or enforce, as many would struggle to identify artificial sentience when it is created. This contrasts with prohibitions on CFCs or human cloning, where violations are comparatively easier to identify.

Secondly, as discussed above, there are numerous potential incentives for creating artificial sentience, including economic and social incentives. Furthermore, if the ban extends to entities which are merely possibly sentient, then this could inhibit the development of advanced AI models which countries may have a strong interest to develop. As such, it is possible that many countries would be unwilling to commit to a global ban (Dung, 2026, p. 157).[27]

A global ban on creating artificial sentience may also be difficult to enforce. The ban may simply be ignored by rogue agents or states in cases of conflict. This risk is greatly exacerbated if space is colonized, as the huge distance between stars may make interstellar enforcement extremely difficult (Torres, 2018).[28]

There are also some reasons for thinking that a global ban today may have backfire risks.[29] For instance, a global ban on artificial sentience may involve effectively banning further research on artificial sentience. This could lead to humans having less understanding of artificial sentience, including knowledge about where it might arise. This could be problematic, as artificial suffering could go unnoticed without this knowledge (though see section 3.5 below for backfire risks associated with increased understanding).

A more speculative worry is that a ban on artificial sentience could lead to less concern for artificial minds. Just as there are no large movements geared towards improving the welfare of clones, there may be limited interest in the welfare of artificial minds if none of them actually exist. Granted, if a ban successfully prohibits the creation of artificial minds then there will be no artificial minds to worry about. But if the ban falls apart or is suddenly lifted, then a general lack of concern toward artificial minds could lead to suffering for those artificial minds created soon after (though see section 3.6 below for backfire risks associated with increased concern for artificial minds).

---

[27] One reviewer pointed out that some classical utilitarians may also be opposed to a permanent ban on the creation of artificial sentience, as such a ban could reduce the total amount of wellbeing that is possible to create.
[28] Though see Delaney and Taylor (2026) for ideas on how interstellar enforcement can be achieved.
[29] See Caviola and Saad (2025) for additional backfire risks that could occur with a ban.

None of these problems seem insurmountable, however. More research can be conducted to understand what kind of ban and enforcement mechanisms are likely to be feasible and accepted by countries and the wider public. Furthermore, if a total ban is likely to inhibit understanding of artificial sentience, then it may be best to limit the ban only to the creation of artificial sentience for commercial reasons.[30] With such a modification, research on artificial consciousness or sentience could continue provided it is for the benefit of artificial minds (or at the very least does not negatively impact them). Finally, it may be possible to increase concern towards artificial minds even if none currently exist today. If these issues are addressed, then it may be possible to implement a robust ban of some kind on the creation of artificial sentience.

Recommended Research Questions

- When should a potential ban extend to (e.g. 2035, 2050, etc)? Under what conditions should the ban be lifted or extended further?
- How extensive should the ban be? Should it completely prohibit the creation of artificial sentience, or just the creation of artificial sentience for "non-essential" reasons?
- What level of uncertainty about artificial sentience is sufficient to justify a prohibition on creating certain AI systems?
- What institutional arrangements would be required to monitor compliance with a ban on creating artificial sentience?
- What lessons can existing international technology bans (e.g. CFCs, human cloning, etc) provide for regulating artificial sentience?

3.2 Ban or Limit the Infliction of Artificial Suffering

Rather than banning the creation of artificial sentience altogether, a ban could instead target the infliction of artificial suffering (Dung, 2026, pp. 166–171). This may be a more appealing strategy if the incentives for creating artificial sentience are too strong, such that a complete ban on creating artificial minds is unlikely to work. Furthermore, it is plausible that a ban on the infliction of

[30] One concern with this is that states or AI companies may continue to create artificial minds for commercial reasons, but simply re-label such work as "research" or some other non-commercial label.

artificial suffering is more likely to be accepted. Most people, for instance, agree that farm animals should not suffer, without necessarily believing that animal agriculture should be abolished altogether (Fonseca & Sanchez-Sabate, 2022; Rosado, 2026). Similarly, many people may come to oppose artificial suffering, without necessarily believing that the creation of artificial sentience should be banned.

A lot of work needs to be done to flesh out this proposal. For instance, it is unclear how strict such a ban should be. If there is only a small chance that some artificial entity is sentient, and a particular action only has a small chance of inflicting a small amount of suffering, then prohibiting such an action may be overbearing and impossible to enforce. A more plausible approach may be to focus just on prohibiting actions which have a high chance of causing significant amounts of artificial suffering.

However, a ban on artificial suffering is susceptible to many of the same problems outlined above in regard to banning artificial sentience. If the ban imposes significant costs (Dung, 2026, p. 167), then some groups will not agree to it. Furthermore, the ban may be difficult to enforce, especially when it comes to rogue actors or in cases of conflict. For this reason, further research needs to be conducted to understand (a) what kind of ban is likely to be accepted, and (b) what kind of enforcement mechanisms are likely to work.

A ban on artificial suffering also requires clearly defined cases of cruelty or suffering that are identifiable for standard regulation. This may be difficult to achieve if it is unclear what qualifies as artificial suffering. This problem is compounded by potential epistemic difficulties when it comes to understanding artificial minds which may be very different from our own.

Relatedly, a ban on artificial suffering may also be susceptible to problems of interpretation. To see this, consider the case of animal welfare laws. Although many countries have anti-cruelty laws (Brels, 2025), what counts as "cruel" can be interpreted in different ways. For some, a fearful animal trapped in a barren cage is not treated cruelly if such treatment is deemed "necessary" for efficient production. A similar problem could occur with artificial suffering. Even if people accept that it's wrong to inflict suffering on artificial minds, they may disagree about what causes artificial suffering, or when an artificial mind may be experiencing suffering.

In order to be successful, a ban on inflicting artificial suffering must provide clear and unambiguous guidance, rather than relying on ambiguous

terminology that is open to interpretation.[31] In addition, institutions could be established to monitor artificial suffering and report on whether the ban is being adhered to. This could be an international regulatory body, similar to the International Atomic Energy Agency (IAEA) (Dung, 2026, p. 162).

Recommended Research Questions

- Is a pervasive ban on the infliction of all artificial suffering feasible? Would it be more feasible to restrict the ban so that it only prohibits the infliction of severe suffering?
- What lessons can be drawn from existing regulations involving nonhuman animals? To what extent have anti-cruelty and similar laws been effective at reducing animal suffering, and would similar effects occur in the context of artificial suffering?
- How should policymakers regulate artificial suffering given uncertainty about artificial sentience? What evidence is needed to prove that artificial suffering has been inflicted in a particular case?
- How could a ban on artificial suffering be monitored and enforced internationally? Would an international oversight agency analogous to the IAEA be feasible?
- How should artificial suffering be defined for legal and regulatory purposes? How can laws and regulations be worded to remove potential ambiguities?

3.3 Identify and Implement Non-Sentient Alternatives

Incentivizing agents not to create artificial minds or inflict artificial suffering may be difficult to do if there are no alternative means of meeting those incentives. For instance, suppose some AI lab is incentivized to create the most powerful and reliable AI model it can. The evidence suggests that the architecture and design most able to achieve this also carries a risk of making the AI sentient. If there are no other alternatives, the AI lab may be strongly

[31] It must, however, still be accurate. One concern is that the ban may focus on things which are easy to measure, rather than things that actually matter. Similar concerns have been raised about AI tools used to monitor animal welfare on farms (Tuyttens et al., 2022).

incentivized to push ahead with this design, even if there is a chance it could create an artificial mind capable of suffering.

Examples such as this suggest that artificial suffering could be reduced by identifying and implementing alternatives that carry less risk of increasing artificial suffering. For instance, in the example above, more research could be conducted to assess whether an equally reliable AI model can be built using an architecture and design that is less likely to result in artificial sentience (Dung, 2026, p. 151).[32] Another example concerns brain organoids. Scientists may have less incentive to create potentially sentient brain organoids if they are able to obtain the sought-after knowledge more easily using simulations of brain organoids. Assuming that digital computer simulations of brain organoids are less likely to be sentient than actual brain organoids, identifying appropriate *in silico* experiments could help to reduce the need to create brain organoids, which in turn could decrease artificial suffering.

This kind of approach can also be found in attempts to reduce animal suffering. There is some evidence to suggest that providing alternatives to animal products (such as vegetarian meals in a cafeteria) decreases meat consumption (Garnett et al., 2019). This idea largely underpins the motivation for creating cultivated animal products such as cultivated meat (otherwise known as "cell-based", "in vitro", "cultured", or "lab-grown" meat; Yu et al., 2025), which aims to replicate conventional meat through cell and tissue culture (Post et al., 2020). If these products are a cheap and delicious alternative to ordinary animal products, then there may be less incentive for individuals to continue consuming those products. Understanding how alternatives have helped to reduce animal suffering could provide some guidance on using alternatives to reduce artificial suffering.

[32] One risk, noted by a reviewer, is that since we know so little about sentience, the very act of creating more kinds of AIs may just raise the possibility of creating sentient AI, even when we are trying to minimise the chance of any given AI being sentient.

It may be impossible to identify alternatives to every incentive regarding artificial sentience. For instance, as some have argued, it may be the case that the more cognitively capable an AI model is, the more likely it is to be sentient (Dung, 2026, p. 152). In which case, it may be difficult to remove any incentives an AI lab might have for creating a powerful model that happens to be sentient. More research needs to be conducted, therefore, to understand what incentives can and cannot be met via alternative options.

Recommended Research Questions

- Which potential alternatives could effectively substitute for creating artificial sentience? What resources are needed to ensure that these alternatives are available?
- How can governments and other actors incentivise the development of alternatives to creating artificial sentience?
- Can AI systems be designed to minimise the probability of artificial sentience while maintaining performance? What AI architectures are least likely to generate artificial sentience?

3.4 Represent Artificial Minds in Political Systems

Another potential strategy to reduce artificial suffering involves including artificial minds in our political systems. This can be done directly, e.g. by giving artificial minds a right to vote, or indirectly, e.g. by having humans represent artificial minds in legislative processes. This kind of political involvement may help to ensure that the interests of artificial minds (either now or in the future) are taken into consideration alongside the interests of humans (Moret et al., n.d.). Similar ideas around political representation have been suggested for groups such as future generations (Eckersley, 2004; Jones et al., 2018) and non-human animals (Cochrane, 2018; Donaldson & Kymlicka, 2011).

If artificial minds or their representatives have a say in the political system, then this may lead to a reduction of artificial suffering. For instance, suppose a state is deciding whether or not to implement some kind of ban on inflicting artificial suffering. Artificial minds (or their representatives) with political representation could make their case as to why such a ban should be implemented, raise concerns that may not be obvious or otherwise considered, and vote accordingly. It may be possible to do this even in situations where artificial minds do not yet exist, yet may do so in the future. In such cases, a

human representative for future artificial minds could vote in a way that aligns with the (predicted) interests of these minds.

Although some political representation for artificial minds could lead to less artificial suffering, implementing such a proposal may be extremely difficult. For one, it may be difficult to determine the interests of artificial minds, especially if they do not yet exist and their interests widely diverge. Even if they did exist, they may not know their own interests well enough, or they might express preferences that lead to more suffering. For instance, they may advocate for there to be no limits on the reproduction of artificial minds, leading to a massive population increase. This could lead to an increase in suffering overall, assuming that a greater number of artificial minds increases the chance that at least some will suffer.

Second, if each (potential) artificial mind is given voting power equivalent to each human then, assuming artificial minds could exist in huge numbers in the future, their interests will dominate any political vote (Shulman & Bostrom, 2021, p. 319). That is, human interests may count for little in the voting process, even if future human generations are also represented. Maybe this is not, in itself, an implausible implication (assuming that artificial minds do not have vastly misaligned values). After all, if artificial minds deserve just as much moral consideration as humans and have sufficient capacities for moral reflection, then it may be unfair to restrict their voting power. But equal voting power would be extremely difficult for most current humans to accept. As such, some forms of political representation for artificial minds are unlikely to be feasible at this stage.

Furthermore, there are major backfire risks with this strategy that need to be taken into consideration. Granting artificial minds political representation could quickly lead to human disempowerment, which may result in human extinction. Additionally, if artificial minds can be easily copied or created, then actors (including humans and AIs) could easily create large numbers of artificial minds to vote according to the actors' interests. This could have significant risks, including risks of totalitarian AI takeover.

For these reasons, it is unclear whether representing artificial minds in political systems is likely to be robustly positive. More research is needed to understand the risks this strategy poses, and identify forms of representation that are more likely to work and be accepted by the public. Additional challenges, such as understanding what interests artificial minds might have in the future, also need to be addressed.

Recommended Research Questions

- How might different forms of political representation lead to a reduction in artificial suffering? What kind of political representation is likely to be the most effective?
- Would it be feasible to give artificial minds equal voting rights? How would this be impacted by expected future populations of artificial minds? Should all artificial minds be given voting rights, or just those that have certain cognitive capacities?
- Should there be any kind of political representation for artificial minds that do not yet exist? If so, how can we predict their interests, and who should represent them? How can we ensure that representatives actually represent the interests of artificial minds?
- To what extent would the public support political representation for artificial minds? How might public support impact what kind of political representation is possible at this stage?
- How should artificial minds be individuated, and what implications might this have for political representation? Should copies of artificial minds be granted equal representation?
- What backfire risks might increased representation have? How can these risks be mitigated?

3.5 Increase our Understanding of Artificial Sentience

One reason as to why humans throughout history have often neglected the interests of nonhumans may be due to a simple lack of understanding. While most humans have an intuitive grasp of when other humans are suffering, it may be difficult to comprehend when some nonhuman animal is suffering. This is especially the case for animals which appear very different from us, such as chickens, fish, or insects. As such, humans may regularly harm these animals without them even knowing. A similar problem could occur with artificial minds. If we do not know when they are suffering, then we may continue to act in ways that cause them considerable harm. As such, in order to address this problem, some have suggested that we need to increase our understanding of artificial suffering (Dung, 2025, pp. 2286–2293; Saad & Bradley, 2025).

For instance, more research could be conducted to know whether there are any indicators in artificial entities (such as certain behaviours or underlying structures) that suggest they may be conscious or sentient. As some have argued, scientific theories of consciousness can be used to empirically test whether some system displays signs indicative of consciousness (Butlin et al., 2026; Dung, 2026, pp. 36–99). For instance, according to the *global workspace theory*, consciousness involves the global broadcast of information to multiple neurocognitive modules, allowing coordination and cooperation between them (Butlin et al., 2026, p. 490; Dehaene & Naccache, 2001). As David Chalmers puts it, the global workspace is essentially "a central clearing-house in the brain for gathering information from numerous non-conscious modules and making information accessible to them. Whatever gets into the global workspace is conscious" (Chalmers, 2023). If an AI system has this feature of global broadcasting, such that information is made available in some kind of workspace to all of its modules, then this serves as a potential indicator of consciousness (Butlin et al., 2026, p. 493).[33] Research along these lines could help to identify which artificial entities are potentially conscious, and thus potentially sentient.

Research can also be conducted to ascertain whether certain behaviour or other aspects are indicative of particular welfare states in artificial entities (Dung, 2026, p. 148). Just as animal welfare science aims to understand when animals have good or bad welfare, a scientific field could be developed geared towards understanding when artificial minds have good or bad welfare. Indeed, some research has already been conducted along these lines (Han et al., 2026; Tagliabue & Dung, 2025).

AI labs could have a large role to play in conducting such research. Anthropic, for instance, has announced that they have started a research programme to understand model welfare (Anthropic, 2025), a move which other AI labs could replicate. AI organisations could also commit to allocating some fraction of their resources to researching AI welfare, just as they have committed to doing for AI alignment (Finnveden, 2024).

This research may be difficult to do however. Relying on behavioural indicators of consciousness, for instance, may not provide an accurate picture of which systems are conscious and which are not. For instance, an LLM may communicate in ways that are indicative of consciousness, but it may just be doing this because of its training data. Being aware of these risks, and designing tests to avoid interfering with the training process in ways that confound

[33] For other scientific theories of consciousness, see Seth and Bayne (2022).

inferences regarding welfare-relevant properties,[34] could be a key part of conducting effective research in this field.

It is plausible that an increased understanding of artificial sentience could help to mitigate incidental risks to artificial minds, as this increased understanding would allow us to detect suffering that might otherwise go undetected. Yet it is important to bear in mind that increased understanding alone is unlikely to be sufficient for reducing such risks. By analogy, even though our understanding of animal welfare has increased significantly over the last century, billions of animals still experience immense amounts of suffering as a result of human practices.[35] An increased understanding of artificial sentience may therefore have little impact unless there is sufficient concern for artificial minds (3.6). This suggests that it may be better to focus on addressing this lack of willingness to reduce suffering, rather than increasing our understanding of artificial minds.

Another major problem with this strategy is that it comes with significant risks (Vinding, 2022b). Increased understanding of artificial consciousness or sentience could exacerbate agential risks such as sadism by giving agents a better understanding of how to increase artificial suffering. As such, it may be necessary to limit who has access to this research, much in the same way as it may be necessary to limit who has access to information on how to create pathogens or other deadly weapons.[36] If this is not possible, or if the risks are large enough, then it may be preferable to prioritize other strategies which are more likely to be robustly positive.

Conducting research on artificial sentience also needs to be done in a way that does not cause suffering in the process. Attempts to understand animal psychology and physiology (often done to increase our understanding of human psychology and physiology) have often involved extremely painful and distressing experiments (Singer, 2015). Similar risks could occur when it comes to research on artificial minds. For this reason, clear guidelines need to be established on how to conduct such research in a way that minimises harm to the subjects.

---

[34] For an example of how this could work, see Susan Schneider's ACT test (Schneider, 2019, pp. 51–57).

[35] However, it is plausible that this number would be even higher in a counterfactual world where significant progress in understanding animal welfare was not made, assuming that an increased understanding of animal welfare motivates at least some people to oppose such practices.

[36] One reviewer commented that such a policy may be implausible, given the difficulty of restricting this research and the potential downsides involved.

Recommended Research Questions

- What knowledge about artificial sentience would be the most valuable to have? To what extent is this knowledge accessible? What tools and other resources are needed to acquire this knowledge?
- What would a science of artificial welfare look like? How can researchers identify positive and negative welfare states in artificial minds?
- Which backfire risks might increased understanding of artificial sentience have? Would limiting who has access to this knowledge minimise such risks? How can this restriction of knowledge be effectively implemented?
- What guidelines can be implemented to ensure that artificial minds do not suffer in the process of studying artificial sentience?

3.6 Increase Concern for Artificial Minds

Another proposed strategy to reduce artificial suffering is to increase concern for artificial minds (Anthis & Paez, 2021). As philosophers such as Peter Singer have argued, humans throughout history have displayed limited concern towards beings outside their "moral circle" – a circle of beings that individuals or society believe deserve (varying degrees of) moral consideration (Figure 6) (Singer, 1981). Individuals may consider their family and friends to deserve the greatest level of moral consideration, followed by members of their tribe, followed by all other humans. Throughout history, some groups of people were considered to be outside the moral circle,[37] such as slaves or people from different ethnicities. Over time, however, the moral circle (of most individuals) has expanded to include all humans. For some, this moral circle has also expanded to include nonhumans such as dogs, cats, and other sentient animals.

[37] Going forward, I use the term "*the* moral circle" to refer to what we may think of as human society's moral circle – aka the circle of beings that humans typically consider to have varying degrees of moral consideration. In the past, *the* moral circle was relatively small, but has since expanded. An individual's moral circle may be different from society's moral circle.

Likewise, we can aim to expand the moral circle to include all sentient beings, including artificial minds (Sebo, 2025).[38]

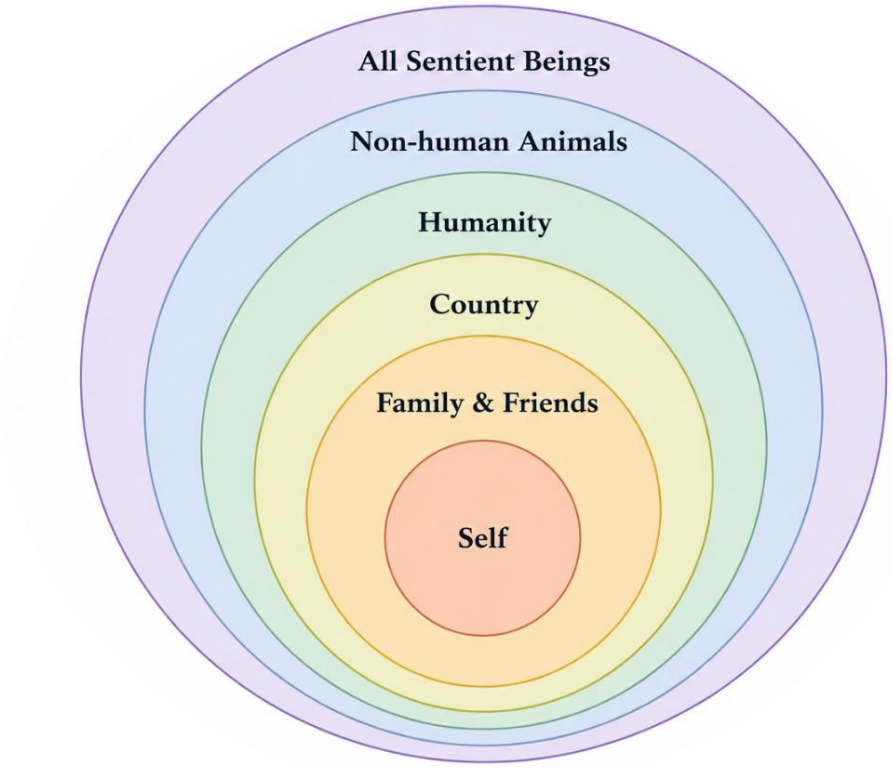

Figure 6. The Moral Circle. Source: ChatGPT

History suggests that when beings are included in the moral circle, they are less likely to suffer (Anthis & Paez, 2021, p. 4). Increased concern for nonhuman animals, for instance, has arguably led to greater public support for improved animal welfare laws which aim to reduce suffering. Greater concern for artificial minds could have a similar effect, as it would increase public support for measures aimed at reducing artificial suffering. Greater concern for artificial minds may also stimulate more research on the topic, as more people would likely be motivated to learn more about how to mitigate artificial suffering.

Other attempts at moral circle expansion suggest that expanding the moral circle to include artificial minds may be difficult to do, however. While

[38] Moral circles could also be expanded to include non-sentient beings and entities, such as ecosystems and the Earth (Brennan & Lo, 2024; Frantz et al., 2025; Leopold, 1949).

significant progress has been made to increase moral concern for both humans and animals, it has not been sufficient. For instance, billions of animals continue to suffer on factory farms (Singer, 2015), despite efforts over the last 50 years by animal advocates to increase concern for their welfare.

These attempts have fallen short for a number of reasons. One reason is that humans tend to empathize less with many non-human animals, especially those very different from humans such as chickens or fish (Miralles et al., 2019; Ragusa, 2025; Suñol et al., 2026). In which case, expanding the moral circle to include artificial minds may be even more difficult due to the potentially huge differences between artificial minds and biological minds.[39] That is, if artificial minds are radically different from other biological minds, this could inhibit understanding and empathy. Furthermore, many people may find it difficult to believe that artificial entities could be sentient, which could likewise inhibit concern.

Some artificial minds in the future may be embodied or have the capacity to communicate with other agents, which could allow for greater understanding and concern. However, it is also possible that many, if not most, artificial minds will lack this capacity for communication. These considerations suggest that increasing our understanding of artificial minds (3.5) may be necessary for effective moral circle expansion.

Strong economic incentives also inhibit significant progress toward moral circle expansion in the case of animals. People may be motivated not to feel moral concern for animals if doing so means they have to give up things which they value, such as cheap animal products. Indeed, there is some evidence to suggest that consuming animal products decreases concern for animals (Loughnan et al., 2010). Similar incentives may exist when it comes to artificial sentience. For instance, increased concern for artificial minds may result in more societal resources being distributed away from humans and towards benefiting artificial minds, or in laws requiring artificial minds to receive fair compensation for their work. Many humans may be motivated not to feel moral concern for artificial minds if doing so means they have to give up resources or access to cheap labour.

Moral circle expansion could also backfire in ways that lead to more artificial suffering (Tomasik, 2015b; Vinding, 2018). For instance, increased concern for

[39] Bostrom suggests that even space alien minds may be easier to understand than artificial minds (Bostrom, 2012, pp. 72–73). This is because aliens, unlike many artificial entities, may be biological creatures that have arisen from evolutionary processes, and thus may share some interests with biological beings on Earth. It should be noted however that Bostrom uses the term "artificial mind" here to refer to artificial intelligence, not conscious minds *per se*.

artificial minds may not translate into less suffering for artificial minds if this concern is misdirected in some sense. To illustrate, consider how increased concern for wild animals may not lead to better wild animal welfare. Individuals that are sympathetic towards wild animals may want to spread ecosystems to other planets, thinking that this would be better for particular species or individual animals. However, as stated above, many individual wild animals suffer considerably as a result of predation, starvation, and disease. In cases such as these, increased concern for beings may not lead to a genuine reduction in suffering.

One example of how this might play out with artificial minds is if people advocate for artificial minds not to be destroyed or switched off, without fully comprehending what this might involve.[40] While some artificial minds may benefit from not being switched off, others may suffer considerably as a result. For instance, some artificial minds may have net negative lives, dominated primarily by feelings of pain and suffering rather than happiness. Yet these minds may be unable to communicate this to others, who may mistakenly think that these minds have positive lives that should be preserved. Consider by analogy pet owners who keep their pets alive out of strong feelings of attachment, even if the animal experiences considerable suffering as a result of some untreatable disease.

If these backfire risks are large enough, then it may be best to avoid moral circle expansion altogether.[41] Alternatively, it may be possible to expand the moral circle in a way that mitigates potential risks.[42] This could involve promoting concern for suffering alongside consideration for artificial minds, as doing so may reduce the likelihood of taking risks which could lead to artificial suffering (Tomasik, 2015b; Vinding, 2018; 2020, pp. 226–238). A lot of uncertainty remains about the feasibility and robustness of these strategies, however. Further

[40] The chance of this occurring may be very small. As one reviewer commented, moral circle expansion would typically be packaged in more robust notions of morality than something simple like "Don't switch off". Still, it is worth flagging that the moral circle could expand in unintended ways.

[41] It is possible that increased concern towards artificial minds may be inevitable. For instance, many people are currently attached to chatbots, e.g. in a romantic capacity (Djufril et al., 2025), or have expressed beliefs that chatbots are conscious (Colombatto & Fleming, 2024). As these chatbots improve (or if they gain other features that could increase empathy and understanding, e.g. a realistic avatar users can interact with on their screens), more people may come to use them and form similar attachments. It is plausible that many of these AIs will become ubiquitous in our lives, acting as our indispensable personal assistants and companions. This could lead to greater concern for (potential) artificial minds, even if nothing is done directly to increase such concern. Concern may also increase if artificial minds advocate for it, or advocate for moral and legal protection (Caviola & Saad, 2025).

[42] This does not entail that such risks are never worth taking. Moral circle expansion may be net positive, even if it carries some risks.

research on moral circle expansion risks, alongside potential strategies for reducing these risks, may be highly valuable at this stage.

Recommended Research Questions

- Would widespread moral advocacy for artificial minds be a good strategy to take at this point? If so, how could this be achieved in a way that is robust to downside risks?
- What lessons can be drawn from attempts to increase concern towards other nonhuman entities (such as nonhuman animals)? Is there any reason to think that increasing concern towards artificial minds will be easier or more difficult?
- To what extent might concern towards artificial minds be limited to a subset of them (e.g. those that can communicate and may be easier to understand)? How can concern for other artificial minds (e.g. those that can't communicate) be increased?
- What backfire risks could occur with unrestricted moral circle expansion? How can these risks be mitigated? Would some form of "controlled" moral circle expansion be optimal, and if so, what would this look like in practice?

3.7 General Strategies for Reducing Suffering

The strategies described above are mainly targeted at reducing artificial suffering in particular. Other strategies have been proposed however which are not focused on reducing artificial suffering *per se*, but on reducing suffering in general. This section provides a brief overview of these proposals, showing how they could help to reduce artificial suffering.[43]

Several strategies for reducing suffering are geared towards preventing conflict. This is because conflict is a risk factor that makes s-risks (especially agential s-risks) more likely (Baumann, 2022, pp. 53–54). Hostility and conflict not only make cooperation on reducing s-risks more difficult, but (as wars throughout history suggest) also increases the chances of large-scale suffering

[43] For more details about these and other strategies, see Baumann (2022) and Vinding (2020).

due to hatred, conflict, and sadism. Reducing conflict, therefore, could be a promising means of reducing artificial suffering.

One strategy for reducing conflicts and agential risks in particular is to reduce malevolent traits in agents. Some s-risks are likely to be amplified by the presence of personality traits such as sadism, psychopathy, and narcissism (Althaus & Baumann, 2020; Baumann, 2022, pp. 54–56). Historical examples of individuals which had these traits, such as Hitler or Stalin, give support to this idea. If malevolent traits in leaders increase the risk of conflict between groups, then reducing these traits could help mitigate s-risks that are more likely to arise in cases of conflict. Reducing traits such as sadism and psychopathy can also help to reduce the risk of individuals inflicting suffering on artificial minds for sadistic reasons. Figuring out how to reduce these traits, therefore, could be a good means of reducing artificial suffering. Some form of moral enhancement is one possible option for achieving this goal (Douglas, 2008; Persson & Savulescu, 2012).

Another approach for mitigating s-risks is to conduct more research on bargaining and conflict escalation. While agential conflict today is initiated solely by human agents (or groups consisting of humans), conflicts in the future could arise primarily from AI systems failing to cooperate. Research in the field of cooperative AI (which aims to equip AI systems with techniques and methods that would allow them to reach mutually beneficial outcomes when interacting with other agents), could therefore be a promising approach to reducing conflict related s-risks (Baumann, 2022, p. 81; Dafoe et al., 2021; DiGiovanni, 2023).

Conflicts could also be reduced by improving our political systems (Vinding, 2022a). For instance, excessive political polarization can lead to hostility and conflict between groups, leading to less cooperation on suffering related issues. Indeed, as some have suggested, artificial minds may become a politicized topic in the future (Caviola & Saad, 2025) (Figure 7).[44] In which case, shifting towards political systems that reduce polarization could help to reduce hostility and conflict that may negatively impact artificial minds. Moving from a presidential system of government to a parliamentary system is one institutional reform that could potentially reduce polarization and other risks that could lead to more suffering (Baumann, 2022; Vinding, 2022a, pp. 232–235). Alternatively, pushing for norms of political tolerance and cooperation may be

[44] See Rost (2026) for examples of state laws declaring that AI systems cannot possess consciousness or moral status. Religion may also play a large part in informing how people think about artificial minds (see e.g. Leo XIV (2026)).

more realistic to achieve, and could have a significant impact on reducing conflict (Baumann, 2021).

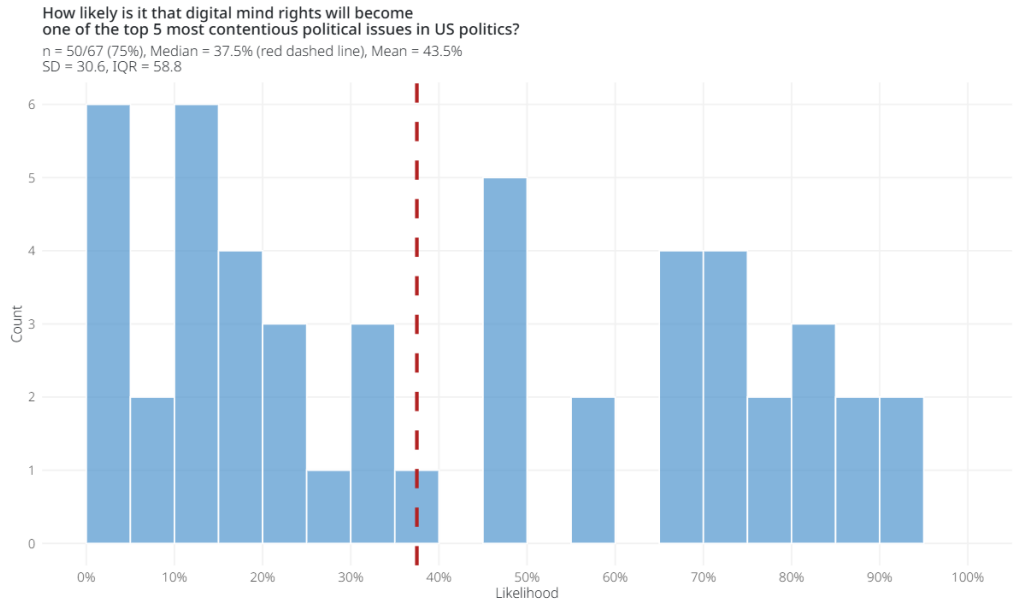

Figure 7. Survey results suggesting that digital minds may be a politicized topic in the future. Survey participants included digital minds researchers, forecasters, AI research and policy experts, and other academic specialists. Source: Caviola and Saad (2025).

Another potential strategy is to ensure that advanced AI systems are aligned with human values (DiGiovanni, 2023). Even if all humans care about artificial minds and want to reduce their suffering, artificial suffering may still occur if we are unable to get AI agents to share these values. An unaligned superintelligent AI, for instance, may create huge amounts of artificial suffering in pursuit of some goal (e.g. if it spawns sentient-subagents to perform certain tasks).[45] By solving the technical problem of alignment, we are better able to shape the values of AI agents in ways that could lead to less artificial suffering. Furthermore, we could use their advanced cognitive capacities to directly work on alleviating artificial suffering (Dung, 2026, p. 187). At present, however, "human values" are not necessarily conducive toward reducing suffering, and may not include concern for artificial minds. As such, it is unclear to what extent this alignment would be beneficial for

[45] This does not mean that only an unaligned superintelligent AI would do this, as artificial suffering could result from an aligned AI following human instructions.

artificial minds. Aligning AI systems with benevolent values that include concern for artificial minds may be preferable.

Space governance may also play a large part in mitigating artificial suffering. As some have argued, the vast distances between stars may make cooperation between interstellar societies difficult (Deudney, 2020; Peppiatt, 2026; Torres, 2018). Colonizing planets in other star systems could therefore increase the risk of conflict, which as stated above may be conducive to greater artificial suffering. It is also possible that incentives for colonizing other planets may create greater incentives for creating artificial minds (Peppiatt, 2026). This is because interstellar travel may be difficult to achieve with biological minds (Marchal et al., 2024; Nunn et al., 2025). If having more incentives for creating artificial minds also increases the risk of artificial suffering (given that more minds would likely be created), then reducing incentives such as these could be beneficial. These considerations suggest that avoiding space colonization could reduce artificial suffering.[46] Alternatively, if it is possible to lock in values that are conducive to reducing suffering prior to space colonization (MacAskill, 2022, pp. 75–101), then risks relating to value drift or interstellar conflict could be reduced.[47] However, this approach carries additional risks as it could lock in values that play out in unexpected ways that may be problematic.

Another strategy is to limit the distribution of technological capacities in a way that reduces the potential for inflicting large amounts of suffering (Baumann, 2022, p. 83). The more individuals that are able to create artificial suffering, the higher the risk. In which case, it may be best to reduce the capacities of any single individual to produce artificial suffering. This could involve, for instance, getting compute manufacturers to somehow block users from creating artificial sentience or artificial suffering. Decisions on how to use powerful AIs could also be limited to decisions made at a collective level, arrived at via democratic processes. By distributing technological capacities in this way, it reduces the chances of any single individual inflicting large amounts of suffering on artificial minds.

3.8 The Strategic Implications of AGI Timelines

What strategies to take may depend heavily on when Artificial General Intelligence (AGI) or Transformative AI (TAI) will arrive.[48] Some have suggested

[46] Though see Ćirković (2019).
[47] Though see Torres (2018, p. 83) for arguments as to why this may fail.
[48] Some scholars prefer to use the term "transformative AI" instead of "AGI" as sufficiently powerful AI could still have a transformative effect on society without necessarily being as intelligent as humans at most cognitive tasks (the necessary requirement for AGI)

that AGI is not likely to arrive for many decades, while others have suggested that it will arrive in just a few years (Amodei, 2026; Dilmegani & Ermut, 2026; Kokotajlo et al., 2025). The arrival of such technology would likely accelerate scientific and technological development, and shape the world dramatically (Amodei, 2024; Bostrom, 2014; Gruetzemacher & Whittlestone, 2022; MacAskill & Moorhouse, 2025).

There are several reasons why the timing of AGI and TAI's arrival may influence what strategies to take regarding artificial suffering. First, if such technology accelerates scientific development, then artificial minds may be more feasible to create. In which case, we may end up with artificial minds in a relatively short period of time if AGI timelines are short.[49] This suggests that, in order to mitigate risks from artificial suffering, we can't rely solely on strategies that take a long time to have an impact.

For instance, increasing concern for artificial minds may take a long time, just as expanding the moral circle to include animals has taken a long time. Some forms of moral circle expansion, such as social media advocacy for AI rights, may therefore not be ideal for reducing artificial suffering in the short term. Other strategies however, such as pushing for state-wide moratoriums on the creation of artificial sentience or limiting access to compute to reduce the number of states or labs capable of creating artificial minds, may be more feasible and likely to have an impact in the short term. Thus, if AGI timelines are short, it may be best to focus on moratoriums and other measures to disincentivize the creation of artificial sentience in the near future. It is important to note that this does not entail abandoning strategies which may take longer. Rather, the claim is simply that we shouldn't rely only on longer term strategies given the potential for AGI or TAI in the near future.

If, on the other hand, AGI timelines are long (such that it is only likely to come about at least a decade from now), then more resources could be allocated towards strategies that are more likely to be successful, even if they take a long time to implement. For instance, if further research reveals that moral circle expansion can be done in a way that avoids backfire risks, and that such a strategy is likely to be extremely effective at reducing suffering, then it could be prioritized, even if it takes lots of time and resources. Ultimately, however,

---

(Gruetzemacher & Whittlestone, 2022). Imagine, for instance, an AI that is capable of making hugely impactful scientific breakthroughs, but can't complete a captcha that most humans can do.

[49] Estimates for when artificial minds might arrive are typically later than estimates for when AGI might arrive. Some reasons for this include that the first AGI systems may lack consciousness, that it may take time to harness energy to do everything we want (including creating artificial minds), and that we may just decide not to create artificial minds (Caviola & Saad, 2025).

we are uncertain about when AGI or TAI will arrive. Because of this, utilising a variety of strategies to mitigate artificial suffering, or prioritizing research and capacity building, may be optimal for now, at least until we have a better understanding of when such technology becomes available.

3.9 Strategies for the Artificial Minds Community

This report has demonstrated that there is a great amount of uncertainty when it comes to artificial minds. It is unclear when artificial sentience might arise, what sources of suffering we should be most concerned about, and what strategies should be taken to reduce this suffering. This section has shown, in particular, that the most common strategies discussed in the literature all face their own set of problems in terms of both feasibility and robustness.

Given this uncertainty, this report recommends that those in the artificial minds community should proceed with caution. Certain actions taken now to reduce artificial suffering could have a massive, irreversible impact on the trajectory of artificial minds. While it is possible that some of these actions will be hugely beneficial, there is a significant risk that some will lead to astronomical amounts of suffering, despite our best intentions. As such, it is crucial to understand these risks, and come to a much more informed position on what strategies are actually worth pursuing.

To reach this better informed position, those in the artificial minds community should seek to understand the full range of views held by those in the community. In particular, members should aim to understand the concerns raised by one another about the robustness of various strategies that aim to reduce artificial suffering. Individuals can work together to identify the main points of disagreement, and collaborate with each other in an attempt to address them (e.g. by sharing research documents across the spectrum of the community for others to give feedback on). This kind of collaboration may be necessary to reduce uncertainty and ensure that any strategies pursued to reduce artificial suffering are robustly positive for artificial minds.

Conclusion

This report has argued that the topic of artificial suffering deserves serious consideration. Many experts agree that some kind of artificial sentience may be possible, and that there is a non-negligible chance of creating it this century. Given the number of artificial minds that could exist in the future, and the amount of suffering they could endure, the moral stakes of the issue are enormous. Ensuring that these lives are not filled with suffering is a moral priority, alongside other pressing issues such as extreme poverty and factory farming.

Despite the multiple avenues by which artificial suffering might be created, these pathways are not inevitable. As this report has shown, there are numerous strategies that could be taken to mitigate these risks. At present, however, it is unclear what strategies should be pursued, largely due to concerns regarding the feasibility and robustness of each strategy. Further work needs to be done to increase the feasibility of promising strategies, reduce potential backfire risks, and identify other potential strategies for reducing artificial suffering.

References

Althaus, D., & Baumann, T. (2020). *Reducing long-term risks from malevolent actors*. Center on Long-Term Risk. https://longtermrisk.org/files/Reducing_long_term_risks_from_malevolent_actors.pdf

Amodei, D. (2024). *Machines of Loving Grace: How AI Could Transform the World for the Better*. darioamodei.com. https://darioamodei.com/essay/machines-of-loving-grace

Amodei, D. (2026). *The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI*. darioamodei.com. https://darioamodei.com/essay/the-adolescence-of-technology

Anthis, J. R., & Paez, E. (2021). Moral circle expansion: A promising strategy to impact the far future. *Futures, 130*, 102756. https://doi.org/10.1016/j.futures.2021.102756

Anthropic. (2025). *Exploring model welfare*. https://www.anthropic.com/research/exploring-model-welfare

Arvan, M., & Maley, C. J. (2022). Panpsychism and AI consciousness. *Synthese, 200*(3), 244. https://doi.org/10.1007/s11229-022-03695-x

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences, 115*(25), 6506–6511. https://doi.org/10.1073/pnas.1711842115

Baumann, T. (2021). *How can we reduce s-risks?* Center for Reducing Suffering. https://centerforreducingsuffering.org/research/how-can-we-reduce-s-risks/

Baumann, T. (2022). *Avoiding The Worst: How to Prevent a Moral Catastrophe*. https://centerforreducingsuffering.org/wp-content/uploads/2022/10/Avoiding_The_Worst_final.pdf

Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines, 22*(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bradley, A., & Saad, B. (2025). AI Alignment Versus AI Ethical Treatment: 10 Challenges. *Analytic Philosophy*. https://doi.org/10.1111/phib.12380

Brels, S. (2025). Non-Cruelty to Animals: A General Principle of (Animal) Law. *LEOH - Journal of Animal Law, Ethics and One Health*, 100–110. https://doi.org/10.58590/leoh.2025.009

Brennan, A., & Lo, N. Y. S. (2024). Environmental Ethics. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/ethics-environmental/

Brooker C. (Writer), & Tibbetts C. (Director). (2014, Dec 16). White Christmas (Season 2, Episode 4). In C. Brooker & A. Jones (Executive Producers), *Black Mirror*. Zeppotron. Netflix.

Browning, H., & Birch, J. (2022). Animal sentience. *Philosophy Compass, 17*(5), e12822. https://doi.org/10.1111/phc3.12822

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences, 30*(6), 488–501. https://doi.org/10.1016/j.tics.2025.10.011

Caviola, L., & Saad, B. (2025). Futures with Digital Minds: Expert Forecasts in 2025. *arXiv: 2508.00536*. https://doi.org/10.48550/arXiv.2508.00536
Cepelewicz, J. (2020). An Ethical Future for Brain Organoids Takes Shape. *Quanta Magazine*. https://www.quantamagazine.org/an-ethical-future-for-brain-organoids-takes-shape-20200123/
Chalmers, D. (1995a). Absent Qualia, Fading Qualia, Dancing Qualia. In T. Metzinger (Ed.), *Conscious Experience*. Imprint Academic.
Chalmers, D. (1995b). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.
Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies, 17*(9-10), 7–65.
Chalmers, D. (2023). *Could a Large Language Model Be Conscious?* Boston Review. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/
Chalmers, D. J. (2025). What we talk to when we talk to language models. https://philarchive.org/rec/CHAWWT-8
Ćirković, M. M. (2019). Space colonization remains the only long-term option for humanity: A reply to Torres. *Futures, 105*, 166–173. https://doi.org/10.1016/j.futures.2018.09.006
Cochrane, A. (2018). *Sentientist Politics: A Theory of Global Inter-Species Justice*. Oxford University Press.
Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness, 2024*(1), niae013. https://doi.org/10.1093/nc/niae013
Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: machines must learn to find common ground. *Nature, 593*, 33–36. https://doi.org/10.1038/d41586-021-01170-0
Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486–492. https://doi.org/10.1126/science.aan8871
Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*(1), 1–37. https://doi.org/https://doi.org/10.1016/S0010-0277(00)00123-2
Delaney, O., & Taylor, M. (2026). *A nightwatchman on every probe: Superintelligent surveillance to prevent galactic anarchy*. Forethought. https://www.forethought.org/research/nightwatchmen
Deudney, D. (2020). *Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity*. Oxford University Press.
DiGiovanni, A. (2023). *Beginner's guide to reducing s-risks*. Center on Long-Term Risk. https://longtermrisk.org/beginners-guide-to-reducing-s-risks/
Dilmegani, C., & Ermut, S. (2026). *AGI/Singularity: 9,800 Predictions Analyzed*. AIMultiple. https://aimultiple.com/artificial-general-intelligence-singularity-timing
Djufril, R., Frampton, J. R., & Knobloch-Westerwick, S. (2025). Love, marriage, pregnancy: Commitment processes in romantic relationships with AI chatbots. *Computers in Human Behavior: Artificial Humans, 4*, 100155. https://doi.org/10.1016/j.chbah.2025.100155
Donaldson, S., & Kymlicka, W. (2011). *Zoopolis: A Political Theory of Animal Rights*. Oxford University Press.
Douglas, T. (2008). Moral Enhancement. *Journal of Applied Philosophy, 25*(3), 228–245. https://doi.org/10.1111/j.1468-5930.2008.00412.x
Dung, L. (2025). How to deal with risks of AI suffering. *Inquiry, 68*(7), 2281–2309. https://doi.org/10.1080/0020174X.2023.2238287

Dung, L. (2026). *Saving Artificial Minds: Understanding and Preventing AI Suffering*. Routledge.
Dung, L. (forthcoming). Why I am not a biological naturalist. *Behavioral and Brain Sciences*. https://philarchive.org/archive/DUNWIA-4
Eckersley, R. (2004). *The Green State: Rethinking Democracy and Sovereignty*. MIT Press.
Farisco, M., Evers, K., & Changeux, J.-P. (2024). Is artificial consciousness achievable? Lessons from the human brain. *Neural Networks*, *180*, 106714. https://doi.org/10.1016/j.neunet.2024.106714
Finnveden, L. (2024). *Project Ideas: Sentience and Rights of Digital Minds*. Forethought. https://www.forethought.org/research/project-ideas-sentience-and-rights-of-digital-minds
Fitch, W. T., Allen, C., & Roskies, A. L. (2025). The evolutionary functions of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *380*(1939). https://doi.org/10.1098/rstb.2024.0299
Fonseca, R. P., & Sanchez-Sabate, R. (2022). Consumers' Attitudes towards Animal Suffering: A Systematic Review on Awareness, Willingness and Dietary Change. *International Journal of Environmental Research and Public Health*, *19*(23), 16372. https://doi.org/10.3390/ijerph192316372
Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neuroscience of Consciousness*, *2022*(1), niac011. https://doi.org/10.1093/nc/niac011
Frantz, P., Rego, F., & Barbas, S. (2025). Ecocentrism vs. Anthropocentrism: To the Core of the Dilemma to Overcome It. *The Linacre Quarterly*, *92*(4), 449–459. https://doi.org/10.1177/00243639251339844
Garnett, E. E., Balmford, A., Sandbrook, C., Pilling, M. A., & Marteau, T. M. (2019). Impact of increasing vegetarian availability on meal selection and sales in cafeterias. *Proceedings of the National Academy of Sciences*, *116*(42), 20923–20929. https://doi.org/10.1073/pnas.1907207116
GenV. (2020). https://genv.org/wp-content/uploads/2020/06/pigs-in-factory-farm.jpg
Gill, S. S., Cetinkaya, O., Marrone, S., Claudino, D., Haunschild, D., Schlote, L., Wu, H., Ottaviani, C., Liu, X., Machupalli, S. P., Kaur, K., Arora, P., Liu, J., Farouk, A., Song, H. H., Uhlig, S., & Ramamohanarao, K. (2024). *Quantum Computing: Vision and Challenges*. arXiv:2403.02240. https://arxiv.org/abs/2403.02240
Godfrey-Smith, P. (2016). Mind, Matter, and Metabolism. *The Journal of Philosophy*, *113*(10), 481–506.
Gonier, D., Adduci, A., & LoCascio, C. (2023). *Cross Fertilizing Empathy from Brain to Machine as a Value Alignment Strategy*. arXiv:2312.07579. https://doi.org/10.48550/arXiv.2312.07579
Gruetzemacher, R., & Whittlestone, J. (2022). The transformative potential of artificial intelligence. *Futures*, *135*, 102884. https://doi.org/10.1016/j.futures.2021.102884
Han, A. Q., Chalmers, D. J., & Izmailov, P. (2026). *How's it going? Reinforcement learning in language models recruits a functional welfare axis*. arXiv preprint. https://arxiv.org/abs/2605.30232
Ho, J. Q. H., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. *Computers in Human Behavior Reports*, *19*, 100715. https://doi.org/10.1016/j.chbr.2025.100715
Hollanek, T., & Nowaczyk-Basińska, K. (2024). Griefbots, Deadbots, Postmortem Avatars: on Responsible Applications of Generative AI in the Digital Afterlife Industry. *Philosophy & Technology*, *37*(2), 63. https://doi.org/10.1007/s13347-024-00744-w

Hongxi, W., Ruting, W., Yiyang, L., Qinglian, H., Jibing, C., & Feng, J. (2025). Human Brain Organoids: Development and Applications. *J Microbiol Biotechnol, 35*, e2411040. https://doi.org/10.4014/jmb.2411.11040

Horta, O. (2015). The Problem of Evil in Nature: Evolutionary Bases of the Prevalence of Disvalue. *Relations, 3*(1), 17–32. https://doi.org/10.7358/rela-2015-001-hort

Jeziorski, J., Brandt, R., Evans, J. H., Campana, W., Kalichman, M., Thompson, E., Goldstein, L., Koch, C., & Muotri, A. R. (2023). Brain organoids, consciousness, ethics and moral status. *Seminars in Cell & Developmental Biology*, *144*, 97–102. https://doi.org/10.1016/j.semcdb.2022.03.020

Johannsen, K. (2021). *Wild Animal Ethics: The Moral and Political Problem of Wild Animal Suffering*. Routledge.

Jones, N., O'Brien, M., & Ryan, T. (2018). Representation of future generations in United Kingdom policy-making. *Futures*, *102*, 153–163. https://doi.org/10.1016/j.futures.2018.01.007

Knutsson, S. (2022). Is anything more pleasant than freedom from unpleasantness? https://www.simonknutsson.com/undisturbedness-as-the-hedonic-ceiling/

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025). *AI 2027*. AI Futures Project. https://ai-2027.com/

Koplin, J. J., & Savulescu, J. (2019). Moral Limits of Brain Organoid Research. *Journal of Law, Medicine & Ethics*, *47*(4), 760–767. https://doi.org/10.1177/1073110519897789

Kudithipudi, D., Schuman, C., Vineyard, C. M., Pandit, T., Merkel, C., Kubendran, R., Aimone, J. B., Orchard, G., Mayr, C., Benosman, R., Hays, J., Young, C., Bartolozzi, C., Majumdar, A., Cardwell, S. G., Payvand, M., Buckley, S., Kulkarni, S., Gonzalez, H. A.,...Furber, S. (2025). Neuromorphic computing at scale. *Nature, 637*(8047), 801–812. https://doi.org/10.1038/s41586-024-08253-8

Kuhn, R. L. (2024). A landscape of consciousness: Toward a taxonomy of explanations and implications. *Progress in Biophysics and Molecular Biology*, *190*, 28–169. https://doi.org/10.1016/j.pbiomolbio.2023.12.003

Lacalli, T. (2024). The function(s) of consciousness: an evolutionary perspective. *Frontiers in Psychology, 15*. https://doi.org/10.3389/fpsyg.2024.1493423

Langlois, A. (2017). The global governance of human cloning: the case of UNESCO. *Palgrave Communications, 3*(1), 17019. https://doi.org/10.1057/palcomms.2017.19

Leo XIV. (2026). *Magnifica Humanitas*. Vatican. https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html

Leopold, A. (1949). *A Sand County Almanac and Sketches Here and There*. Oxford University Press.

Long, R., Sebo, J., & Sims, T. (2025). Is there a tension between AI safety and AI welfare? *Philosophical Studies, 182*(7), 2005–2033. https://doi.org/10.1007/s11098-025-02302-2

Loughnan, S., Haslam, N., & Bastian, B. (2010). The role of meat consumption in the denial of moral status and mind to meat animals. *Appetite*, *55*(1), 156–159. https://doi.org/https://doi.org/10.1016/j.appet.2010.05.043

MacAskill, W. (2022). *What We Owe the Future: A Million-Year View*. Oneworld.

MacAskill, W., Meissner, D., & Chappell, R. Y. (2023). Elements and Types of Utilitarianism. In R. Y. Chappell, D. Meissner, & W. MacAskill (Eds.), *An Introduction to Utilitarianism*. Utilitarianism.net. https://utilitarianism.net/types-of-utilitarianism/

MacAskill, W., & Moorhouse, F. (2025). *Preparing for the Intelligence Explosion*. Forethought. https://www.forethought.org/research/preparing-for-the-intelligence-explosion

Mamak, K. (2026). In defense of artificial suffering. *Philosophical Studies*. https://doi.org/10.1007/s11098-026-02493-2

Marchal, S., Choukér, A., Bereiter-Hahn, J., Kraus, A., Grimm, D., & Krüger, M. (2024). Challenges for the human immune system after leaving Earth. *npj Microgravity, 10*(1), 106. https://doi.org/10.1038/s41526-024-00446-9

Mayerfield, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

McIntyre, J. H. (2026). Individuating Artificial Minds. *Erkenntnis*. https://doi.org/10.1007/s10670-026-01097-w

McMahan, J. (2015). The Moral Problem of Predation. In A. Chignell, T. Cuneo, & M. C. Halteman (Eds.), *Philosophy Comes to Dinner: Arguments About the Ethics of Eating* (pp. 268–293). Routledge. https://doi.org/10.4324/9780203154410

Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness, 8*(1), 43–66. https://doi.org/10.1142/s270507852150003x

Milinkovic, B., & Aru, J. (2026). On biological and artificial consciousness: A case for biological computationalism. *Neuroscience & Biobehavioral Reviews, 181*, 106524. https://doi.org/10.1016/j.neubiorev.2025.106524

Miller, C. A., & Bennett, I. (2008). Thinking longer term about technology: is there value in science fiction-inspired approaches to constructing futures? *Science and Public Policy, 35*(8), 597–606. https://doi.org/10.3152/030234208X370666

Miller, W. B., Baluška, F., Reber, A. S., & Slijepčević, P. (2025). Biological mechanisms contradict AI consciousness: The spaces between the notes. *BioSystems, 247*, 105387. https://doi.org/10.1016/j.biosystems.2024.105387

Miralles, A., Raymond, M., & Lecointre, G. (2019). Empathy and compassion toward other species decrease with evolutionary divergence time. *Scientific Reports, 9*(1), 19555. https://doi.org/10.1038/s41598-019-56006-9

Mogensen, A. L. (2025). How to resist the Fading Qualia Argument. *Synthese, 206*(5), 252. https://doi.org/10.1007/s11229-025-05338-3

Moret, A. (2025). AI welfare risks. *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02343-7

Moret, A., Magaña, P., & Paez, E. (n.d.). *The Political Claims of Sentient Beings to AI Governance*. [Working paper].

Mullally, N. (2026). The self-preservation test for artificial sentience. *AI and Ethics, 6*(1), 142. https://doi.org/10.1007/s43681-026-00983-x

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review, 83*(4), 435–450. https://doi.org/10.2307/2183914

Neven, H., Zalcman, A., Read, P., Kosik, K. S., van der Molen, T., Bouwmeester, D., Bodnia, E., Turin, L., & Koch, C. (2024). Testing the Conjecture That Quantum Processes Create Conscious Experience. *Entropy, 26*(6), 460.

Newen, A., & Montemayor, C. (2025). Three types of phenomenal consciousness and their functional roles: unfolding the ALARM theory of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences, 380*(1939), 20240314. https://doi.org/10.1098/rstb.2024.0314

Ng, Y.-K. (1995). Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering. *Biology & Philosophy, 10*(3), 255–285. https://doi.org/10.1007/BF00852469

Niikawa, T., Hayashi, Y., Shepherd, J., & Sawai, T. (2022). Human Brain Organoids and Consciousness. *Neuroethics, 15*(1), 5. https://doi.org/10.1007/s12152-022-09483-1

Nunn, A., Bell, J., & Guy, G. (2025). Are we trapped on Earth: space and accelerated ageing. *npj Microgravity, 12*(1), 10. https://doi.org/10.1038/s41526-025-00551-3

Pauketat, J. (2021). The Terminology of Artificial Sentience. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/sujwf
Pauketat, J. V. T., Ladak, A., & Anthis, J. R. (2024). World-making for a future with sentient AI. *British Journal of Social Psychology, 64*(1), e12844. https://doi.org/10.1111/bjso.12844
Peppiatt, C. (2026). *The Case Against the Stars*. Self-Published. https://philpapers.org/rec/PEPTCA
Persson, I., & Savulescu, J. (2012). *Unfit for the Future: The Need for Moral Enhancement*. Oxford University Press.
PhilPapers. (2020). *Survey Results: Consciousness: identity theory, panpsychism, dualism, functionalism, or eliminativism?* PhilPapers. https://survey2020.philpeople.org/survey/results/5010
Porębski, A., & Figura, J. (2025). There is no such thing as conscious artificial intelligence. *Humanities and Social Sciences Communications, 12*(1), 1647. https://doi.org/10.1057/s41599-025-05868-8
Post, M. J., Levenberg, S., Kaplan, D. L., Genovese, N., Fu, J., Bryant, C. J., Negowetti, N., Verzijden, K., & Moutsatsou, P. (2020). Scientific, sustainability and regulatory challenges of cultured meat. *Nature Food, 1*(7), 403–415. https://doi.org/10.1038/s43016-020-0112-z
Ragusa, A. (2025). Why Humans Prefer Phylogenetically Closer Species: An Evolutionary, Neurocognitive, and Cultural Synthesis. *Biology, 14*(10), 1438. https://www.mdpi.com/2079-7737/14/10/1438
Ratner, P. (2018). Matrioshka Brain: How advanced civilizations could reshape reality. *Big Think*. https://bigthink.com/hard-science/are-we-living-inside-a-matrioshka-brain-how-advanced-civilizations-could-reshape-reality/
Register, C. (2025). Individuating artificial moral patients. *Philosophical Studies, 182*(11), 3225–3246. https://doi.org/10.1007/s11098-025-02409-6
Ritchie, H., Rodés-Guirao, L., Mathieu, E., Mersmann, S., Bachler, D., Dattani, S., Gerber, M., Ortiz-Ospina, E., Hasell, J., & Roser, M. (2023). *Population Growth*. Our World in Data. https://ourworldindata.org/population-growth
Rosado, P. (2026). *Most people care about farm animals — our food system doesn't reflect that*. Our World in Data. https://ourworldindata.org/most-people-care-about-farm-animals-our-food-system-doesnt-reflect-that#article-citation
Rost, T. (2026). Legislating AI Consciousness Without an Exit. *The Regulatory Review*. https://www.theregreview.org/2026/06/29/rost-legislating-ai-consciousness-without-an-exit/
Roumbanis, L. (2026). Suffering machines? A critical inquiry concerning the moral risks of creating an 'organoid intelligence'. *Futures, 176*, 103758. https://doi.org/10.1016/j.futures.2025.103758
Saad, B. (2026). *On Anil Seth's "Conscious artificial intelligence and biological naturalism"*. Meditations on Digital Minds. https://meditationsondigitalminds.substack.com/p/on-anil-seths-conscious-artificial
Saad, B., & Bradley, A. (2025). Digital suffering: why it's a problem and how to prevent it. *Inquiry, 68*(7), 2110–2145. https://doi.org/10.1080/0020174X.2022.2144442
Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*. Princeton University Press.
Schukraft, J. (2020). *The subjective experience of time: welfare implications*. Rethink Priorities. https://rethinkpriorities.org/research-area/the-subjective-experience-of-time-welfare-implications/
Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 459–479). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0017

Searle, J. (2007). Biological Naturalism. In M. Velmans & S. Schneider (Eds.), *The Blackwell Companion to Consciousness* (pp. 325–334). Blackwell.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756
Sebo, J. (2025). *The moral circle. Who matters, what matters, and why*. W.W. Norton & Company.
Sebo, J., & Long, R. (2025). Moral consideration for AI systems by 2030. *AI and Ethics, 5*(1), 591–606. https://doi.org/10.1007/s43681-023-00379-1
Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1–42. https://doi.org/10.1017/S0140525X25000032
Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience, 23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4
Shulman, C., & Bostrom, N. (2021). Sharing the World with Digital Minds. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status* (pp. 306–326). Oxford University Press. https://doi.org/10.1093/oso/9780192894076.003.0018
Singer, P. (1981). *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press.
Singer, P. (2015). *Animal Liberation* (40th Anniversary ed.). Open Road Media.
Suñol, M., Bastian, B., & López-Solà, M. (2026). Empathy for pain in humans and animals differs based on species, psychosocial and cultural factors. *Scientific Reports, 16*(1), 9605. https://doi.org/10.1038/s41598-025-32047-1
Tagliabue, V., & Dung, L. (2025). *Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare*. arXiv preprint. https://arxiv.org/abs/2509.07961
Thagard, P. (2023). Cognitive Science. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2023/entries/cognitive-science/
Tomasik, B. (2014). *Do Artificial Reinforcement-Learning Agents Matter Morally?* arXiv:1410.8233. https://doi.org/10.48550/arXiv.1410.8233
Tomasik, B. (2015a). *Artificial Intelligence and Its Implications for Future Suffering*. Center on Long-Term Risk. https://longtermrisk.org/artificial-intelligence-and-its-implications-for-future-suffering
Tomasik, B. (2015b). *Reasons to Promote Suffering-Focused Ethics*. Reducing Suffering. https://reducing-suffering.org/the-case-for-promoting-suffering-focused-ethics/
Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167
Torres, P. (2018). Space colonization and suffering risks: Reassessing the "maxipok rule". *Futures, 100*, 74–85. https://doi.org/10.1016/j.futures.2018.04.008
Tuyttens, F. A. M., Molento, C. F. M., & Benaissa, S. (2022). Twelve Threats of Precision Livestock Farming (PLF) for Animal Welfare [Review]. *Frontiers in Veterinary Science, 9*. https://doi.org/10.3389/fvets.2022.889623
Vinding, M. (2018). *Moral Circle Expansion Might Increase Future Suffering*. https://magnusvinding.com/2018/09/04/moral-circle-expansion-might-increase-future-suffering/
Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. https://magnusvinding.com/wp-content/uploads/2020/05/suffering-focused-ethics.pdf
Vinding, M. (2022a). *Reasoned Politics*. Ratio Ethica. https://magnusvinding.com/wp-content/uploads/2022/03/reasoned-politics.pdf
Vinding, M. (2022b). *Why I don't prioritize consciousness research*. https://magnusvinding.com/2022/07/07/why-i-dont-prioritize-consciousness-research/

Vinding, M. (2026). *Defusing cluelessness with degrees of justification*. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/upnKGdTMMDvcLamMh/defusing-cluelessness-with-degrees-of-justification

Višak, T. (2022). *Capacity for Welfare across Species*. Oxford University Press.

Young, P. J., Harper, A. B., Huntingford, C., Paul, N. D., Morgenstern, O., Newman, P. A., Oman, L. D., Madronich, S., & Garcia, R. R. (2021). The Montreal Protocol protects the terrestrial carbon sink. *Nature, 596*(7872), 384–388. https://doi.org/10.1038/s41586-021-03737-3

Yu, Y., Wassmann, B., Lanz, M., & Siegrist, M. (2025). Willingness to consume cultured meat: A meta-analysis. *Trends in Food Science & Technology*, *164*, 105226. https://doi.org/10.1016/j.tifs.2025.105226