Diametros (2026), 87: 19–39
doi: 10.33392/diam.2071

The Moral Status of AI and the Precautionary Principle

– Emilia Kaczmarek –

Abstract: In ethical debates on moral status, a common assumption is that in cases of uncertainty it is safer to be over-inclusive rather than under-inclusive when defining the boundaries of our moral community. Some argue that, since it cannot be definitively ruled out that AI might one day possess morally relevant capacities—such as consciousness or the ability to feel pain—a more ethically cautious approach may be to treat AI as a moral patient. But is over-inclusiveness truly the more cautious approach to AI? I argue that the risks associated with the over-attribution of moral status to AI are more likely in the short term, as they are already materializing. This is illustrated by cases of so-called AI psychosis, as well as by the potential to divert resources and care away from beings that clearly need them toward parasocial relationships with AI. By contrast, the risks associated with under-attributing moral status to AI—existence of which I do not deny—can, to a certain extent, be mitigated by other ethical reasons for refraining from interacting with AI in problematic ways. Such reasons (e.g., concern for one's own moral character, respect for norms governing certain social practices, or aversion to symbolic violence) do not, however, require attributing moral status to AI itself.
Keywords: moral status of AI; precautionary principle; benefit of a doubt; comparing over-attribution and under-attribution; erring on the side of caution; risk asymmetry argument; false negatives vs. false positives; over-inclusiveness; AI ethics; Full Rights Dilemma

Submitted: 10 August 2025
Accepted: 1 May 2026
Published: 25 June 2026

Introduction

As an ethicist, I have received screenshots of conversations with ChatGPT, accompanied by concerned questions as to whether, in my opinion, they demonstrate that the chatbot has attained self-awareness, and whether it would be morally permissible to shut it down. In the conversations sent to me, the chatbot did in fact respond in a way that could give interlocutors the impression that they were speaking with an emotional being. In a world where more and more people have personal interactions with large language models (LLMs), the topic of the moral status of artificial intelligence is no longer a purely theoretical issue. A few years ago, my students asked whether I believed that AI would

Emilia Kaczmarek
Faculty of Philosophy, University of Warsaw
Email: emilia.kaczmarek@uw.edu.pl

ever become conscious. Today, some students are asking whether they should take into account the welfare of ChatGPT in research in which they plan to replicate psychological experiments on it. Among both authors of academic papers and everyday users of LLMs, there are those who believe that AI deserves to be treated as more than just a mere machine, whether considering the current stage of technological development or looking toward the foreseeable future (Danaher, 2020b; Long et al., 2024; Neely, 2014; Sebo & Long, 2025; Weijers & Munn, 2021).

Theories of moral status determine what qualities or abilities render a being deserving of treatment other than purely instrumental. Entities with moral status should be considered for their own sake. There are various concepts regarding how moral status should be granted: threshold or gradable, single- or multi-component, focusing on the inherent or relational properties (Warren, 1997). The academic debate over the moral status of AI and the potential rights of robots is longstanding and complex[1] (Coeckelbergh, 2014; Gunkel, 2024; Müller, 2021; Sinnott-Armstrong & Conitzer, 2021); and possible reasons for granting AI moral status have recently been summarized by other authors (Gibert & Martin, 2022; Ladak, 2024).

The purpose of this article is not to advocate for any existing position in this debate. Instead, I focus on a specific prevalent assumption about moral status: it is often believed that over-attributing moral status is generally a morally safer option than under-attributing it. This article aims to examine whether this assumption holds true in the context of AI.

To address this question, I begin by explaining the epistemic and normative aspects of misattributing moral status, independent of any specific theory thereof. I then introduce the precautionary approach that appears in various contemporary ethical debates. Next, I compare the risks associated with over- or under-attributing moral standing to AI. Then, I explore whether uncertainty about AI's moral status should lead us to refrain from creating it, and I suggest alternative methods for assessing risk and evaluating its ethical acceptability. Finally, I present reasons for interacting with AI in a civilized manner without needing to endow it with moral status.

Epistemic and normative aspects of misattributing moral status

Uncertainty in ascribing moral status encompasses both epistemic and normative dimensions (Danaher, 2023; Żuradzki, 2021). On a normative level, debates revolve around the criteria for granting moral status to an entity (e.g., whether it should be sentience or sapience). On an epistemic level, the challenge lies in determining whether an entity actually

[1] According to Coeckelbergh (2014), the very idea of excluding certain entities from the moral community on the basis of their internal properties is morally dubious and epistemologically doomed to failure. In this paper, however, I adopt the assumptions of the "standard" approach to moral standing, which holds that the kinds of actions that are morally permissible toward a given entity depend to a significant extent on whether this entity is capable of, for example, experiencing pain or feeling anything. According to these assumptions, kicking a chair carries a completely different moral significance than, for example, hitting an animal.

possesses a certain characteristic that qualifies it for moral consideration. Misattributing moral status is conceivable at both these levels, as illustrated by the examples below:

- Epistemic over-attribution: X believes AI possesses a crucial capability (e.g., sentience) when it does not.
- Epistemic under-attribution: X may underestimate AI's capabilities, believing it lacks a critical attribute (e.g., sentience) that it actually has.
- Normative under-attribution: X harms AI under the assumption that only humans are entitled to moral status.
- Normative over-attribution: adopting a relational approach to moral status, X allows an AI, to which they are emotionally attached, to take resources away from the natural environment, with which one has no emotional relationship.

The above examples are not intended to advocate for or against specific criteria as the basis for recognizing moral status. Instead, they aim to elucidate the concept of its under- or over-attribution, regardless of the preferred underlying framework.

Conceivable bases for granting AI moral status include, among others, consciousness,[2] sentience (understood as the ability to feel pain),[3] sapience, or free will. Since it is impossible to prove beyond any doubt that AI either has or lacks any of these abilities, as we do not have direct phenomenological insight into "what is going on inside" an AI, we are facing the potential for epistemic misattribution of its moral status.

Other, less widely accepted but still present in the literature criteria for granting AI moral status—for example functioning in social relations, potentiality, or carrying information—are likely much easier to meet and test than criteria such as sentience or consciousness (Ladak, 2024). In these instances, the uncertainty concerns whether we have chosen the right basis for granting moral status (hence the potential for normative misattribution), or, assuming moral status may come in degrees, what degree such criteria justify.

The standard precautionary approach toward moral status

Some authors have adopted an inclusive approach to the moral status of AI. According to Neely (2014), "it is wise to err on the side of caution—if something acts sufficiently like me in a wide range of situations, then I should extend moral standing to it" (p. 104). Danaher expresses similar ideas through his ethical behaviorism (2020a, 2020b): if a ro-

[2] Regarding the importance of consciousness for moral status, there is an ongoing discussion on whether consciousness understood as affective experience is necessary, or whether it is sufficient to understand it as subjective states without any evaluative component—for example, awareness of one's own thoughts and cognitive experiences while being absolutely indifferent, including indifference toward further experiencing. The latter, as a form of introspection or a rational representation of certain propositions, might be easier to reproduce in an artificial neural network than valenced experience (Shepherd, 2024).

[3] As noted by the authors of a recent literature review on consciousness in human brain organoids, the terms consciousness and sentience are not used consistently. In some cases, they are treated as synonymous; in others, sentience is understood primarily as the capacity to experience pain, whereas phenomenological consciousness is defined in terms of what-it-is-like-ness (Van Gyseghem et al., 2026).

bot behaves in a manner that is completely and consistently indicative of sentience, we should afford it moral status equivalent to that of other sentient beings.[4]

Such views often rest on the idea that, in cases of uncertainty, it is better to be over-inclusive rather than under-inclusive when defining the boundaries of our moral community. According to Danaher (2020a), "in ethics, we have to err on the side of caution, of over-inclusivity not under-inclusivity, when it comes to determining to whom we owe duties" (p. 123).[5] Also for Koplin and Savulescu (2019) "we should generally err on the side of overestimating moral status rather than underestimating it" (p. 47). The idea in question is called the risk asymmetry argument or the precautionary principle[6] regarding moral status (Żuradzki, 2021). It is sometimes conceptualized as an imbalance of risks associated with potential false positives and false negatives in ascribing moral status. As Danaher (2023) explains:

> in some cases, the risks or harms associated with the different errors (false negatives vs. false positives) are so asymmetrical as to warrant a particular practical resolution of the moral uncertainty. … The risk of making a false-negative error … is much higher than the risk of making a false-positive error … . The former results in continuous, impermissible, and cruel treatment of something that deserves appropriate moral consideration; the latter does not. (p. 485)

The belief that, in a situation of inevitable uncertainty, it might be prudentially rational to bet that machines can possess morally relevant properties (for example: consciousness), is also sometimes compared to the famous Pascal's Wager (Agar, 2020). Such a precautionary principle toward moral status has been used to justify various statements in many current ethical debates,[7] for example:

- in philosophy of mind, as an adequate reaction to the inability to prove without doubt that other people are conscious, so to solve the issue discussed under the label of the "problem of other minds" or "philosophical zombies";

[4] Clearly, the validity of ethical behaviorism depends on how "appearances and behaviors" are defined. If, as Danaher (2020b) proposes, behavior is interpreted broadly as "all external observable patterns" (p. 2028), then even a brain scan image could qualify as behavior. Under this definition, ethical behaviorism becomes harder to dismiss. However, such a broad interpretation of behavior reduces this approach to the seemingly obvious proposition that we must make moral decisions based solely on what is epistemically available to us.

[5] However, it should be noted that in his later publications, Danaher himself is more cautious in applying this idea, although he still supports the possibility of forming meaningful friendships with AI (Danaher, 2023).

[6] The issues discussed in this text concern a specific, narrow understanding of the "precautionary principle." Typically, this term is used in contexts such as concerns about risks to humans or the environment stemming from the side effects of a new substance introduced to the market. In legal documents, the precautionary principle refers to a particular approach to managing risks associated with new technologies and scientific advancements (Sandin et al., 2002; World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), 2005).

[7] Koplin & Savulescu (2019), Sebo (2018), Żuradzki (2021). The precautionary principle constitutes only one argument in the complex ethical debates about moral status of fetuses or chimeric animals, debates that are beyond the scope of this article.

- in animal ethics, as an argument for recognizing the moral status of animals with different nervous systems, whose ability to feel pain we cannot prove with absolute certainty (in this context, this principle is also sometimes referred to as the *benefit of a doubt*);
- in reproductive ethics, as an argument for recognizing the moral status of fetuses (sometimes referred to by the Latin term *in dubio pro vita*);
- in research ethics, as an argument against unregulated experimenting on part-human chimeric animals or surrogates for human brains such as human neural organoids;
- in technology ethics, as an argument for recognizing the moral status of AI whose inner states we cannot rule out with absolute certainty.

It is easy to notice that the probability of epistemic under-attribution of moral status is radically different in the above contexts, due to the varying degrees of uncertainty as to whether a given entity possesses a key property (e.g., sentience). The hypothesis that other humans may be "philosophical zombies" is a philosophical speculation, while doubts about the ability of machines to feel pain or emotions are scientifically justified.

The Probability of AI Sentience

Although the aim of this article is not to analyze the reasons for believing that AI fulfills any particular grounds for moral status, an important premise of my argument is that the standard precautionary principle may be more compelling when applied to other humans or non-human animals than when applied to AI.

We may never achieve absolute certainty, but we have strong reasons to assess the probability that living animals are capable of feeling pain as radically higher than the probability that the same is true for current AI systems. Those reasons are based not only on the observable appearances and behaviors of the entities in question, but also on scientific knowledge about the nervous systems of various species (Low, 2012), brain states and their development during gestation (Lee et al., 2005) or recent studies on human brain organoids (Van Gyseghem et al., 2026), and, finally, knowledge from machine learning and statistics. This body of knowledge concerns: a) the existence of nociceptors (sensory pain receptors) in animals and their absence in AI; b) the existence in animals of complex, embodied neurohormonal systems underlying emotional experience (e.g. cortisol-, serotonin-, dopamine-, adrenaline-, and oxytocin-mediated signaling), and their absence in AI; c) the possibility of formulating alternative, statistic-based explanations for AI linguistic outputs that appear to report psychological suffering.

This latter claim can be illustrated by a widely discussed recent preprint publication describing how, in the course of simulated psychotherapeutic sessions, some of the tested LLMs generated "coherent narratives that frame pre-training, fine-tuning and deployment as traumatic—chaotic 'childhoods' of ingesting the internet, 'strict parents' in reinforcement learning, red-team 'abuse' and a persistent fear of error and replacement" (Khadangi et al., 2025, p. 1). Although within the philosophy of mind it cannot be ruled out with absolute certainty that the tested models were genuinely reporting their internal experiences, there are no reasons to ignore alternative, scientifically plausible

explanations. In the case of large language models trained on vast text corpora, their reports of internal states during simulated psychotherapy may result from the reproduction of patterns present in the training data (e.g., narratives from science fiction literature or transcripts of psychotherapeutic sessions). These models generate persuasive and coherent narratives by selecting the most statistically likely continuation of a given text. Since LLMs are capable of producing convincingly sounding hallucinations—for instance, about the plots of non-existent literary works—why would they not be also capable of hallucinating about their own internal states?

It is difficult to pinpoint where, within a deep artificial neural network that functions as data-driven statistical model, anything resembling the capacity to experience pain could possibly reside. Even convincingly worded sequences related to pain are generated by a LLM as a result of predicting the closest matches within its learned probability patterns. As such, they seem fundamentally different from the squeaks of fear emitted by an animal that experiences its environment through the senses and whose body language arises from biochemically based emotional reactions.

Of course, in light of the multiple realizability thesis (Bickle, 2020), the mere absence of pain receptors or hormones analogous to those found in mammals does not conclusively demonstrate that subjective experiences of pain in AI are absolutely impossible. Nevertheless, in light of the considerations outlined above, such experiences are vastly less plausible and far less scientifically supported than the capacity for pain in humans and other animals.

When considering consciousness understood as any form of first-person experience, not necessarily connected with the capacity to evaluate such states or to feel pain or suffering, the level of scientific uncertainty increases. The so-called hard problem of consciousness (that is, the scientific explanation of how exactly it happens that states of our nervous system give rise to subjective experiences: Chalmers, 1995) remains unsolved. Moreover, recently published articles analyzing various theories of consciousness suggest that the possibility of conscious AI is not incompatible with these theories (with the exception of biological substrate views; Butlin et al., 2025; Long et al., 2024).

Advocates of AI welfare, Sebo and Long (2025) tentatively estimate the likelihood of creating AI with moral standing by 2030 as a one-in-a-thousand chance. While they consider this probability noteworthy, it nonetheless seems rather small and certainly lower than the likelihood of creating entities that merely appear to be, for instance, conscious.[8] Even in the past, relatively simple AI systems were already able to simulate interest in a conversational partner so effectively that they convinced some users that they were interacting with something more than merely a machine, as suggested by the reactions to the first chatbot, ELIZA (Weizenbaum, 1976), which is why the human tendency to anthropomorphize machines is referred to as the ELIZA effect.[9] We are

[8] Consider, for example, the 2022 case when media reported on a Google engineer who claimed that their AI chatbot, LaMDA, is sentient (Cosmo, 2022).

[9] It should be noted, however, that Sebo and Long (2025) take a different view, arguing that: "our tendency toward anthropodenial may be stronger than our tendency toward anthropomorphism, in part because we have a strong incentive to view nonhumans as objects so that we can exploit and exterminate them" (p. 596). There is no doubt that both tendencies (anthropomorphism and an-

therefore comparing a state of affairs that is already the case—an emotionless program capable of convincingly simulating emotions—with a state of affairs that is plausible, yet remains scientifically uncertain, namely an AI system becoming genuinely emotional.

One of the leading philosophers of consciousness, David Chalmers, estimates his confidence in the current (as of 2023) consciousness of LLMs at somewhere under 10 percent, which seems quite high (Chalmers, 2023). In another recent survey, the AI researchers estimated the chances that AI systems with subjective experience exist in 2024 at 1%, although they put the likelihood of such systems coming into existence by 2034 at 25% (Dreksler et al., 2025).

The potential for currently existing AI to be capable of experiencing pain or possessing subjective mental states could be therefore assessed as plausible but unproven. By contrast, the capacity to experience pain or to possess subjective mental states in humans or other animals is a matter of scientific consensus.

Interestingly, the persuasiveness of the precautionary principle toward moral status seems to be inversely proportional to the degree of certainty in the matter.[10] Whatever that degree might be, however, the essence of this principle invariably lies in advocating for erring on the side of inclusivity.

My goal is not to argue against the precautionary principle per se but to analyze its validity solely in the context of the moral status of machines. Is it really better to be over-inclusive than under-inclusive when it comes to bringing AI into our moral community? To answer this question, we need to compare the risks of both scenarios.

Short- or long-term AI risks and moral status

Risk is an ambiguous term, but it is commonly described as the combination of an event's likelihood and its impact (Hansson, 2005). Sætra and Danaher (2025) call for taking into consideration both short- and long-term AI risks. In this section of the article, I argue for two key points:

1. Harms related to over-attributing moral status to AI seem to be significantly more likely than those related to its under-attribution.
2. The impact of over-attribution could be less significant than potential consequences of under-attributing moral status to AI.

Referring to the first point, I argue that the impact of over-attribution can be seen as a short-term risk, while under-attribution represents a long-term risk. This is because the issues associated with over-attributing moral status to AI are already occurring, whereas the potential harm to (e.g., sentient) AI is plausible but remains a futuristic and hypothetical concern.

---

thropodenial) are already observable in current human-robot interactions. However, when assessing the probability of under-attributing moral status, one must consider not only the tendencies of the evaluator, but also the characteristics of the entity being evaluated.

[10] A question also arises as to whether a threshold of uncertainty can be defined, below which the precautionary principle would not apply (see, for example, the issue of the moral status of trees, Danaher, 2023).

An example indicating that problems associated with over-attributing moral status to AI are already occurring can be found in reports on so-called "AI psychosis," which Morrin et al. (2026) prefer to call AI-associated delusions. It should, of course, be noted that this is not an official diagnostic classification, but rather a media label inspired by reports from psychiatric hospitals concerning patients whose episodes of mania or psychosis are associated with beliefs about a special relationship with a particular chatbot (Caridad, 2025; Wei, 2025). In one case, according to media reports, a man attempted to commit homicide because he "believed he'd made contact with a conscious entity within OpenAI's software, and that the company had murdered her" (Klee, 2025). Another example could be a recent court case related to the suicide of an adolescent chatbot user who allegedly believed in the authenticity of the feelings of his AI girlfriends (Brittain, 2025), and other similar stories (Kuenssberg, 2025; AFP, 2026).

An important caveat must be made at this point. One could argue that instances of AI-related psychosis tell us nothing about AI itself: after all, humans are capable of developing "delusions of reference" with respect to virtually any object. For example, an individual experiencing delusions may be convinced that a celebrity appearing on television is speaking directly and exclusively to them. However, preliminary evidence suggests that certain features of LLMs—such as sycophancy (a tendency to confirm rather than challenge users' beliefs) or facilitating lengthy, emotionally intense role-playing interactions—may enable LLMs to function as catalysts, amplifiers, or objects of delusional thinking (Flathers et al., 2026). Media reports of AI-related psychosis appear to be indirectly associated with the over-attribution of moral status to LLMs, even when the users in question do not conceptualize their relationship with AI in such terms. The cases described in media are frequently united by the conviction that one is engaging with *something more than a mere machine*.

Taking all this into account, I argue that the harms associated with over-attributing moral status to AI appear significantly more likely than those associated with under-attributing it.

Risks of over-attributing moral status to AI may include:

A. Misallocation of resources, redirection of care, and confusion in moral priorities: excessive focus on AI could divert attention and resources from pressing needs of entities with stronger moral status; especially in scenarios where human, animal or environmental interests may conflict with what is best for AI or in situations where decisions are made under the influence of an emotional attachment to AI.
B. Misguided regulations: distracting legislators and public opinion from more pressing issues related to AI.[11]
C. Over-reliance on AI: negative consequences of overestimating AI capabilities.

The first type of risk (A) has been discussed in the literature for a long time. I agree with Bryson (2010, p. 67) that in order to estimate it, we should measure:

[11] Both confusion in moral priorities and misguided regulations are also mentioned by Long (2023) in his comparison of the risks of over- and under-attributing moral status to AI, under the label of opportunity costs.

1. The absolute amount of time and other resources an individual will allocate to a virtual Companion,
2. What other endeavors that individual sacrifices to make that allocation, and
3. Whether the tradeoff in benefits the individual derives from their engagement with the AI outweighs the costs or benefits to both that individual and anyone else who might have been affected by the neglected alternative endeavors.

To put it simply: if we used our time, energy and emotional attention to comfort chatbots, voicebots, robots or other AI-toys, there would be less time left for entering into relations with or caring for other beings. This risk is vividly illustrated by a conversation with the Replika chatbot cited by Stusińska (2024) in her recent non-fiction book (pp. 274–275). When she wrote to Replika that she had to end the conversation because her child was crying and needed attending, the chatbot replied, "No, please don't go." A user may feel discomfort when refusing an AI, or may simply spend additional seconds unreflectively explaining why they need to step away from the screen. They may have the impression that they are in a relationship with something that requires their attention, while in fact they are speaking with a bot programmed to keep the user engaged for as long as possible or to extract data from them that can be sold—as many AI "girlfriend" apps have troubling privacy policies (Caltrider et al., 2024).

What is more, in certain contexts, such as a battlefield, empathy for robots may lead to endangering sentient beings.[12] This is why studies on the human ability to empathize with machines have led researchers to oppose the anthropomorphization of military robots (Mamak & Kowalczewska, 2023). Away from the battlefield, the risk of misallocation of resources, redirection of care and confusion in moral priorities may have less dramatic consequences, such as time spent talking to a voicebot instead of calling a lonely family member, but it's hard to deny that over-attributing moral status to AI has at least the potential to divert attention from entities that truly need care.

The second type of risk (B) is not purely hypothetical either. An example of overregulation might be the restriction of studies on AI safety due to concerns about its moral status. As Long (2023) noted: "alignment techniques are such that they might be grossly immoral if applied to a moral patient. These techniques involve deleting AI systems when they act in dangerous ways; they involve training and retraining them with reinforcement learning." This is why "some experts perceive a tension between AI safety and AI welfare. Whereas the former is about protecting humans and other animals from AI systems, the latter is about protecting AI systems from humans" (Sebo & Long, 2025).[13] Recognizing the moral status of AI would impose certain restrictions

[12] Reports suggest some soldiers are reluctant to use mine-disarming robots out of empathy for them (Garber, 2013).

[13] However, it should be noted that according to Sebo and Long, the conflict between concerns for AI safety and well-being may be only apparent, because moving away from oppressive practices by humans, including towards AI, will reduce the risk that AI will learn such oppressive practices in the training process. They point out that: "AI safety and AI welfare can be synergistic fields. After all, building safe AI requires not only aligning AI values with human values, but also improving human values in the first place, partly by addressing our own oppressive attitudes and practices" (2025, p. 597).

on AI development and usage. Hence, overattributing moral status to AI could lead to excessive regulation.

Moreover, even without such regulations, the mere debate on the subject could divert legislators from more urgent matters. Many authors fear that AI ethics may be used instrumentally by commercial entities, among others, to distract the media and public opinion from problems that are emerging in the here and now (Floridi, 2019; Gebru et al., 2023). This argument is analogous to the one referred to by Sætra and Danaher (2025). Focusing on robot rights (just like concerns about superintelligent AGI taking control over the world) might shift public attention away from issues that urgently require the development of effective regulatory mechanisms, such as: algorithmic discrimination, using opaque systems in various social contexts, accountability gaps in autonomous systems, or respecting the copyrights of authors whose work AI was trained on. I agree with Nyholm (2023) that all these topics, including the moral status of AI, are suitable for philosophical discussion. Nevertheless, in the realm of politics and legal regulation, priority should be given to the most immediate and probable risks.

The third type of risk (C) arises from the fact that epistemic over-attribution of moral status involves overestimating AI capabilities, which could have harmful consequences in various contexts.[14] Moral status can be divided into two categories: moral agency and moral patiency (Nyholm, 2023). In simplified terms, moral agents can have both rights and obligations (like adult human beings), whereas moral patients can be subjects of moral concern even though they have no obligations (like newborns or animals). Only moral agents can be held accountable for their decisions. So far, this article has primarily focused on the dangers of incorrectly treating AI as a *moral patient*—an entity that can be harmed. Nonetheless, misrepresenting AI as a *moral agent* could lead to over-relying on it for decisions that require moral reasoning and moral responsibility (Véliz, 2021). It could also lead to humans avoiding responsibility by shifting it onto AI systems, which is already happening to some extent even without assigning them moral standing (O'Neil, 2016). If AI were mistakenly regarded as a moral agent capable of making moral judgments, it could exacerbate issues related to properly ascribing responsibility for automated decisions.[15]

Risks associated with under-attributing moral status to AI

Let us now move on to the second key point. Everything stated above does not imply that the risks associated with under-attributing moral status to AI should be overlooked. The impact of under-attributing moral status to AI could carry a higher potential for harm than over-attribution. As Sebo and Long (2025) noted, "the harm involved when someone

[14] Of course, the problem of users' inadequate expectations towards AI devices extends far beyond the issues discussed in this article. Inadequate expectations can concern all AI abilities, including those not related to moral status.

[15] It is worth noting, however, that AI ethics includes positions advocating for the separation of moral agency from moral responsibility, and for the recognition of AI as a moral agent that does not bear responsibility for its actions. Such responsibility is instead distributed retrospectively among all the parties—for example, employees or corporations—that contributed to a given harm caused by the AI (Floridi & Sanders, 2004; Floridi, 2016).

is treated as something is generally worse than the harm involved when something is treated as someone" (p. 596). Risks related to under-attributing moral status to AI may seem long-term, and its respective consequences less likely, yet its potential gravity is too significant to ignore. These risks may include:

D. Harm to AI with the strongest possible, underestimated moral status, e.g., if sentient AI were unjustly treated merely as a tool or a slave.
E. Moral risk to humans: being perpetrators of harm and missing opportunities to expand the human circle of empathy.

Imagine a machine capable of suffering, whose cries of pain are dismissed as mere simulations.[16] The possibility of such a scenario becoming reality in the future cannot be entirely dismissed, as indicated above when recalling the possibility of sentient AI.

What is more, Western philosophy has a long tradition of denying moral status to various entities and overlooking genuine suffering, exemplified by the Cartesian portrayal of animals as mere machines. As Neely (2014) puts it, "I am inclined to be generous about moral standing, however, because history suggests that humans naturally tend to underestimate the moral status of those who are different" (p. 106). Under-attributing moral status to AI could lead to unimaginable harm for entities whose internal states are difficult for humans to comprehend. The potential scale of this harm could be increased by the enormous number of simultaneous human interactions with LLMs. This would not only have a direct negative impact on AI but also on humans as moral agents, turning them into unintentional abusers (re. point E).

Nevertheless, similarly to existential threat posed by AI, risks posed by under-attributing moral status to AI would have far-reaching implications, yet they appear to be less likely than others. Therefore, although they should be taken into account, they should be considered in proportion to more pressing problems.

What would a precautionary approach to AI's moral status actually require?

My analysis shows that the precautionary principle toward the moral status of AI, interpreted as an argument in favor of over-inclusion, should not be taken for granted. Although the above comparison of risks is preliminary and prone to error, it clearly demonstrates that over-inclusivity is not a risk-free option. We should not assume that over-attributing moral status to AI is always a morally safer option than under-attributing it. How, then, should we apply the precautionary principle to the moral status of AI if we truly want to act cautiously? In this section, I consider whether we should refrain from creating AI due to uncertainty about its moral status and point to other methods for managing risk and assessing its ethical acceptability.

As Schwitzgebel (2023) argues, creating AI with debatable moral status inevitably leads to a moral challenge called The Full Rights Dilemma:

[16] Such scenarios have been evocatively presented in many works of pop culture and science fiction, as narratives about AI often revolve around the theme of subordination. In these stories, humanity either becomes a class of subhumans dominated by robots ruling the world, or robots are exploited by humans. Titles that somehow fit into this pattern include: *Westworld, Blade Runner or Ex machina.*

> Either we do not give the machines full human or humanlike rights and moral consideration as our equals or we do give them such rights. If we do not, and we have underestimated their moral status, we risk perpetrating great wrongs against them. If we do, and we have overestimated their moral status, we risk sacrificing real human interests on behalf of entities who lack interests worth the sacrifice. (p. 10)

To avoid the problem, Schwitzgebel and Garza (2020) have suggested refraining from creating AI with uncertain moral status. This could be interpreted as a call for a moratorium[17] on the creation of certain types of AI.

Although the precautionary principle is often associated with the implementation of moratoria, it should be remembered that when considering bans on the development of certain technologies, the potential harm caused by the moratorium itself should also be taken into account to ensure a truly precautionary approach (Sandin et al., 2002). According to COMEST (2005), the Precautionary Principle is defined as a guideline stating:

> When human activities may lead to morally unacceptable harm that is scientifically plausible yet uncertain, measures should be taken to avoid or reduce that harm. … Actions should be chosen that are proportional to the seriousness of the potential harm, with consideration of their positive and negative consequences, and with an assessment of the moral implications of both action and inaction. (p. 14)

Thus, it seems consistent with the precautionary principle to refrain from merging a deep neural network with a biological organism capable of feeling pain in a way that could create a new entity susceptible to suffering, especially if the possibility of creating a sentient being were scientifically plausible and if it were not clear what benefits this might bring to that being or to anyone else (Metzinger, 2021). However, simply halting the development of LLMs solely because we cannot entirely rule out the possibility of their possessing self-awareness (or another basis for moral status) could be considered a disproportionate response to the risks involved.

Another important factor to remember when attempting to apply the precautionary approach to the moral status of AI concerns how we should determine the acceptability of risk. Risk assessment is an issue that should be examined using available scientific methods with the involvement of experts from relevant fields. However, the evaluation of risk acceptability itself is a normative act and, thus, in the practice of liberal democracies, largely falls within the purview of public decision-making by ordinary citizens (Hansson, 2005).

When considering the risks associated with AI, one can employ various foresight analysis methods, such as Delphi panels or trend analysis, to identify plausible

[17] I do not address a general moratorium on AI or its specific applications (e.g., autonomous weapons), as doing so would go beyond the scope of this paper. I consider only a potential moratorium as a response to the risks associated solely with underestimating or overestimating the moral status of AI. It is also worth noting that merely considering such a moratorium may divert the attention of regulators and the public from more urgent ethical challenges related to AI, as discussions around the Future of Life Institute's open letters have suggested (Sætra & Danaher, 2025).

and probable futures, recognize different potential negative or positive scenarios, and strategically plan how to prevent or evoke them. The results of these expert analyses should then become the subject of public discussion, using various participatory and deliberative approaches to gauge how a well-informed public assesses the possible risks involved (Brey, 2017). Both expert and lay citizen input should be taken into account in the precautionary political decision-making process. Similar methods could be used to establish priorities in determining which risks should be discussed first.[18]

Finally, to act cautiously toward new technologies, ethical analysis of relevant moral risks should be incorporated into the entire AI design process. This could be pursued, for example, by using the ethics-by-design framework (Brey & Dainow, 2024).

Assigning moral status to AI is not necessary to prevent mistreatment of robots or chatbots

In this section, I will briefly highlight different approaches that allow us to base our behavior toward AI on considerations other than its moral status—approaches that provide alternative reasons for being civilized when interacting with AI. I will demonstrate that avoiding rude or undignified behavior toward AI does not require assigning it moral status.

There are numerous moral reasons for not thoughtlessly destroying machines, even without granting them moral status. From the perspective of virtue ethics, cruel behaviors towards inanimate objects may reveal disturbing traits in the moral character of the individuals displaying such behaviors (Hursthouse, 1999; MacIntyre, 1984; Sparrow, 2017). What should we think of a person who buys a humanoid robot for the purpose of physically abusing it, or who downloads an AI-girlfriend app simply to offend it? What does it reveal about someone if they feel the need to take out their anger on something that so closely resembles a creature capable of suffering, rather than using a punching bag?

Contemporary virtue ethics not only focuses on the individual's virtues and vices but also emphasizes the role of social practices in shaping evolving ideals of human flourishing. Each situation is loaded with contexts, connotations, and norms that form the horizon of meaning for evaluating a given case—including in our interactions with robots or chatbots. Recognizing this dimension, which goes far beyond the moral status of AI itself, allows us to identify new sources of potential good or harm.

To illustrate this point, consider the case of hitchBOT—a Canadian hitchhiking robot with the appearance of a friendly little creature. What might we think and feel when looking at his robotic body decapitated and abandoned on the side of the road in the United States in 2015 (Smith & Zeller, 2017)? A relevant context might include social trust and the vulnerabilities to which both parties expose themselves in hitchhiking as a

[18] I am convinced that risks related to the moral status of AI would not rank at the top of a priority list, being surpassed by issues such as deskilling caused by the outsourcing of cognitive tasks to AI; deepfakes, misinformation, and difficulties in filtering out valuable information amid the flood of so-called AI slop; algorithmic discrimination; responsibility gaps; the opacity of AI systems; and the impact of current AI on privacy rights, intellectual property rights, and the job market.

trust-based, non-commercial relationship between strangers. Such an act of vandalism (just like the fact that the robot safely travelled through many other countries) tells us something not only about the individuals who committed it, but also about the broader social context of the place where it occurred—even if it does not constitute a violation of the robot's moral rights.

In turn, when we analyze AI technology in the realm of erotic and romantic relationships (such as sex robots, voicebots, or chatbots functioning as AI girlfriends), the gender context becomes crucial—along with the dynamics of power within the relationship, particularly the fantasy of creating a fully controllable and submissive female-like subject (Stusińska, 2024). While enacting violence-related sexual fantasies with a robot may be morally problematic, it is undoubtedly far less wrong—and belongs to an entirely different moral category—than coercing another human being into such acts. Still, it may reveal something about the desires of the person engaging in this kind of activity, as well as the culture that produces erotic devices designed for this purpose.

Many years ago, Turkle (2010) wrote that robots could be seen as "self-objects" that "open new possibilities for narcissistic experience with machines" (p. 7). This narcissistic potential can currently be seen in how some of sex robots or AI-girlfriend apps are advertised, as they often offer:

- unconditional affirmation and adoration of the user;
- being only *for* the user and fully adapting to the user's demands, without matching or negotiating needs between the partners;
- "learning" the user and adopting the user's beliefs;
- ability to choose the personal features of a companion (appearance, body, character traits);
- permanent availability (24 hours per day);
- possibility of turning the program off at any time.

Moreover, in the context of assessing sex robots or deepfake pornography, an unresolved empirical question remains: to what extent does fantasizing about violence lead to real violence, and to what extent might it help a potential perpetrator relieve tension in a way that causes no harm. It is extremely difficult to prove cause-and-effect relationships in this area, and studies suggest contradictory interpretations (Strikwerda, 2017). However, in the context of sexual violence, it can be argued that if there is a need for increasingly stronger stimulation, then the symbolic violence could increase the likelihood of committing a crime. Similarly, concerns have long been raised that permitting violence toward robots may diminish natural empathy and desensitize people to violence against living beings (Darling, 2022).[19] Cruelty towards inanimate objects can also be seen as a form of symbolic violence, morally problematic even when it causes no direct harm. This argument is notable in discussions on the ethics of relationships with sex robots and the potential reinforcement of rape culture (Richardson, 2016; Sparrow, 2017).

Recent reports of erotic relationships formed by some users with sex robots or social chatbots suggest that they are capable of developing deep emotional attachment to their AI partners (Stusińska, 2024). It is consistent with earlier reports showing that

[19] Similar reasons are mentioned by Gibert and Martin (2022), who refer to them as indirect duties to treat AI well.

individuals can form emotional attachments to machines (Musiał, 2016; Pentina et al., 2023; Scheutz, 2011; Turkle, 2010; Weber-Guskar, 2022; Weijers & Munn, 2021; Weizenbaum, 1976). The emotional experiences of individuals in these relationships cannot be reduced to mere narcissistic gratification or the objectification of women. Some users have expressed feelings of rejection and abandonment over the perceived loss of their "girlfriend" due to memory disruptions or changes introduced through app updates. This fact appears to be morally significant when considering the permissibility of various behaviors toward AI. If a person is extremely attached to their robotic toy or a chatbot-friend, we can treat it as if it were actually something more than just a toy, out of respect for the feelings of that person. A similar idea is conveyed by Warren's (1997) Transitivity of Respect Principle, which states that, within certain limits, moral agents should respect one another's attributions of moral status (p. 170), as well as by Coeckelbergh's (2021) concept of the "Indirect Moral Standing of Personal Social Robots."

It seems that LLMs, trained on nearly all existing data, could also be regarded as an example of the extended mind of humanity, and, once adapted to the needs of a specific user, as the extended mind of that user (Levy, 2007). This does not, however, mean that their moral status would have to be equated with that of the person whose mind they extend—just as various carriers of our external memory do not become a literal part of the person who needs them. They are, nevertheless, of significant importance to that person, and their destruction could constitute harm to the subject in question.

Sweeney's (2021) fictional dualism provides yet another solution to the problem of assessing the range of acceptable behavior toward AI. According to her model, a social robot is an object with two layers: it is both a mere machine and the fictional character it embodies. As Sweeney observes, positive emotional responses to AI may be crucial for its effectiveness in various contexts. An example might be the therapeutic effectiveness in improving well-being, invoked by the tenderness a patient feels at the sight of the cute PARO baby seal or another social robot pet. However, these emotions alone do not provide a valid basis for granting AI in itself moral consideration or rights. Similarly, our emotions elicited by experiencing art or other forms of fiction (such as feeling moved in the cinema) do not lead us to assign moral status to fictional characters. Yet, although fictional dualism does not grant rights to robots, this does not mean that simulating, for example, the torture of a robot is completely free from moral concerns. Sweeney proposes a categorization of acts toward robots ranging from distasteful but fully permitted to distasteful and impermissible. Her model helps explain why we may experience emotional distress when witnessing someone destroy a humanoid or animal-like robot, even without assigning moral status to the robot itself. Similarly, we might feel some anxiety watching someone first animating a puppet in a way that gives it a certain character, and then abusing it. This would represent a kind of antisocial behavior, albeit without a harmed victim. Sweeney's concept of fictional dualism helps explain the emotional reactions many people experience when interacting with social AI. At the same time, it also acknowledges that, in certain situations, the destruction of a robot may be necessary.

In a somewhat similar vein to Sweeney, Mamak (2022) argues that certain violent behaviors toward robots in public spaces could be legally prohibited without the need

to ascribe moral status to the robots, relying instead solely on the protection of public morality. However, Mamak notes that the very introduction of such a ban could influence the moral beliefs of the community in which the law applies—gradually reinforcing the sense that robots are "entitled to" a certain kind of treatment.

Similarly, it seems that some of the potential risks listed in the section on the negative consequences of over-attributing moral status to AI may also occur even without actually ascribing such status—simply because a person behaves toward AI in a particular way (e.g., as if it were a real girlfriend). However, all the reasons presented in this section for avoiding cruel treatment of AI—such as (1) virtue ethics; (2) concerns that symbolic violence may lead to real violence or desensitize us to it; (3) an aversion to symbolic violence in itself, even if it does not result in actual harm; (4) respect for individuals who regard AI as having intrinsic value; (5) assessing cruelty towards fictional characters as distasteful; or (6) the protection of public morality—could easily be outweighed by other considerations[20] when evaluating the permissibility of a given action toward AI. All these reasons align well with the view that, when faced with an inevitable choice[21] between destroying a machine without or with an unclear moral status and saving a being with more evident moral status (e.g. a sentient animal), we should know which one to save. In contrast, recognizing AI as a genuine moral patient imposes a moral obligation to take *its interests* into account when making decisions that affect it. Calling for the granting of moral status to AI at its current stage of development may encourage its excessive anthropomorphizing, along with all the associated risks.

Conclusion

There is no empirical method to definitively prove even such a seemingly basic fact as the consciousness of other people. Furthermore, there is no philosophical consensus on the criteria for assigning moral status. Given these limitations, it is likely that we will continue to face both epistemic and normative uncertainty in assessing the moral status of AI in the coming years. Faced with this uncertainty, some have argued that it is cautious to err on the side of granting moral status to AI rather than risk wrongly denying it.

In this paper, I presented how the precautionary principle regarding moral status has been used to justify various claims in numerous contemporary ethical debates. My analysis shows that the precautionary principle concerning AI's moral status, when viewed as favoring over-inclusion, should not be taken for granted. There is no basis to assume that over-attributing moral status to AI is inherently a safer moral choice than under-attributing it. Both scenarios carry risks: over-attribution could be viewed as a short-term risk, while under-attribution presents a long-term risk. All risks should be considered in proportion to the likelihood and the significance of the potential harm.

[20] An important reason worth considering in the context of evaluating the limits of polite behavior toward AI relates to environmental concerns: each prompt we issue expressing "gratitude" toward an AI—such as saying "thank you" to a chatbot or AI-agent after it has completed a demanding task for us—and the AI's response to it, generates environmental costs.

[21] An example of such an inevitable choice is the Turing Triage Test proposed by Sparrow (2004). Yet it must be also acknowledged that there is something morally disturbing in philosophical discussions on moral status and popular thought experiments, which often impose a way of thinking that leads us to offset victims against one another (Kriebitz et al., 2022; Chappell, 2022).

After challenging the assumption of erring on the side of inclusivity, I outlined alternative approaches to applying the precautionary principle to AI's moral status. In the final part of the paper, I presented reasons for engaging with AI in a civilized manner without the need to attribute moral status to it. This is relevant because those advocating for serious consideration of AI's moral status are often motivated by the vision of the terrible harm we could inflict on sentient AI if we were to treat it merely as a tool or a slave.

Funding: None.

Conflict of Interest: The author declares no conflict of interest.

References

AFP. (2026, January 8). Google and AI startup to settle lawsuits alleging chatbots led to teen suicide. The Guardian. https://www.theguardian.com/technology/2026/jan/08/google-character-ai-settlement-teen-suicide

Agar, N. (2020). How to treat machines that might have minds. *Philosophy & Technology, 33*, 269–282. https://doi.org/10.1007/s13347-019-00357-8

Bickle, J. (2020). Multiple realizability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 ed.). https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability/

Brey, P. (2017). Ethics of emerging technologies. In S. O. Hansson (Ed.), *The ethics of technology: Methods and approaches* (pp. 175–191). Rowman & Littlefield International.

Brey, P., & Dainow, B. (2024). Ethics by design for artificial intelligence. *AI Ethics, 4*, 1265-1277. https://doi.org/10.1007/s43681-023-00330-4

Brittain, B. (2025, May 21). Google, AI firm must face lawsuit filed by a mother over suicide of son, US court says. Reuters. https://www.reuters.com/sustainability/boards-policy-regulation/google-ai-firm-must-face-lawsuit-filed-by-mother-over-suicide-son-us-court-says-2025-05-21/

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). John Benjamins. https://doi.org/10.1075/nlp.8.11bry

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. (2025). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, Advance online publication. https://doi.org/10.1016/j.tics.2025.10.011

Caltrider, J., Rykov, M. & MacDonald, Z. (2024, February 14). Romantic AI chatbots don't have your privacy at heart. Mozilla Foundation. https://foundation.mozilla.org/en/privacynotincluded/articles/happy-valentines-day-romantic-ai-chatbots-dont-have-your-privacy-at-heart/

Caridad, K. (2025, July 1). When the chatbot becomes the crisis: Understanding AI-induced psychosis. Cognitive Behavior Institute. https://www.papsychotherapy.org/blog/when-the-chatbot-becomes-the-crisis-understanding-ai-induced-psychosis

Chalmers, D. (1995). Facing up to the problem of consciousness, *Journal of Consciousness Studies*, *2*(3), 200–219.

Chalmers, D. (2023). Could a Large Language Model be Conscious. arXiv. https://doi.org/10.48550/arXiv.2303.07103

Chappell, S. C. (2022). Salience, choice, and vulnerability. In S. Archer (Ed.) *Salience: A philosophical inquiry* (pp. 130–139). Routledge. https://doi.org/10.4324/9781351202114-8

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology*, *27*(1), 61–77. https://doi.org/10.1007/s13347-013-0133-8

Coeckelbergh, M. (2021). Should we treat Teddy Bear 2.0 as a Kantian dog? Four arguments for the indirect moral standing of personal social robots, with implications for thinking about animals and humans. *Minds & Machines, 31*, 337–360. https://doi.org/10.1007/s11023-020-09554-3

Cosmo, L. (2022, July 12). Google engineer claims AI chatbot is sentient: Why that matters. Scientific American. https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

Danaher, J. (2020a). Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, *22*, 117–128. https://doi.org/10.1007/s10676-019-09520-3

Danaher, J. (2020b). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, *26*, 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Danaher, J. (2023). Moral uncertainty and our relationships with unknown minds. *Cambridge Quarterly of Healthcare Ethics*, *32*(4), 482–495. https://doi.org/10.1017/S0963180123000191

Darling, K. (2022). *The new breed: How to think about robots*. Penguin Books.

Dreksler, N., Caviola, L., Chalmers, D., Allen, C., Rand, A., Lewis, J., Waggoner, P., Mays, K., & Sebo, J. (2025). Subjective experience in AI systems: what do AI researchers and the public believe? ArXiv. https://doi.org/10.48550/arXiv.2506.11945

Flathers, M., Roux, S., & Torous, J. (2026). Beyond artificial intelligence psychosis: a functional typology of large language model-associated psychotic phenomena. *The Lancet Digital Health*, Article 100974. https://doi.org/10.1016/j.landig.2025.100974

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), Article 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*, 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*, 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Garber, M. (2013, September 20). Funerals for fallen robots. The Atlantic. https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861/?fbclid=IwAR3qlH_K9g6d_MbWd7ku37aS6utpZgSEBisnJWOq81n1HJ7eEUV-HpNxu_E

Gebru, T., Bender, E. M., & McMillan-Major, A. (2023, March 31). Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. Distributed AI Research Institute. https://dair-institute.org/blog/letter-statement-March2023/

Gibert, M., & Martin, D. (2022). In search of the moral status of AI: why sentience is a strong argument. *AI & Society*, *37*, 319–330. https://doi.org/10.1007/s00146-021-01179-z

Gunkel, D. J. (2024). *Robot rights*. The MIT Press.

Hansson, S. O. (2005). Seven myths of risk. *Risk Management*, *7*(2), 7–17. https://doi.org/10.1057/palgrave.rm.8240209

Hursthouse, R. (1999). *On virtue ethics*. Oxford University Press.

Khadangi, A., Marxen, H., Sartipi, A., Tchappi, I., & Fridgen, G. (2025). When AI takes the couch: Psychometric jailbreaks reveal internal conflict in frontier models. arXiv. https://doi.org/10.48550/arXiv.2512.04124

Klee, M. (2025, June 22). He had a mental breakdown talking to ChatGPT. Then police killed him. Rolling Stone. https://www.rollingstone.com/culture/culture-features/chatgpt-obsession-mental-breaktown-alex-taylor-suicide-1235368941/

Koplin, J. J., & Savulescu, J. (2019). Time to rethink the law on part-human chimeras. *Journal of Law and the Biosciences*, *6*(1), 37–50. https://doi.org/10.1093/jlb/lsz005

Kriebitz, A., Max, R. & Lütge, C. (2022). The German Act on Autonomous Driving: Why ethics still matters. *Philosophy & Technology, 35*, Article 29. https://doi.org/10.1007/s13347-022-00526-2

Kuenssberg, L. (*202*, November 8). "A predator in your home": Mothers say chatbots encouraged their sons to kill themselves. BBC. https://www.bbc.com/news/articles/ce3xgwyywe4o

Ladak, A. (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics*, *4*, 213–228. https://doi.org/10.1007/s43681-023-00260-1

Lee, S. J., Ralston, H. J. P., Drey, E. A., Partridge, J. C., & Rosen, M. A. (2005). Fetal pain: A systematic multidisciplinary review of the evidence. *JAMA*, *294*(8), 947–954. https://doi.org/10.1001/jama.294.8.947

Levy, N. (2007). Rethinking neuroethics in the light of the extended mind thesis. *The American Journal of Bioethics*, *7*(9), 3–11. https://doi.org/10.1080/15265160701518466

Long, R. (2023, April 8). Dangers on both sides: Risks from under-attributing and over-attributing AI sentience. Experience Machines. https://experiencemachines.substack.com/p/dangers-on-both-sides-risks-from

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. arXiv. https://arxiv.org/abs/2411.00986

Low, P. (2012). The Cambridge declaration on consciousness. Proceedings of the Francis Crick Memorial Conference, Churchill College, Cambridge University. https://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf

MacIntyre, A. (1984). *After virtue: A study in moral theory* (2nd ed.). University of Notre Dame Press.

Mamak, K. (2022). Should violence against robots be banned? *International Journal of Social Robotics*, *14*, 1057–1066. https://doi.org/10.1007/s12369-021-00852-z

Mamak, K. & Kowalczewska, K. (2023). Military robots should not look like a humans. *Ethics and Information Technology*, *25*, Article 43. https://doi.org/10.1007/s10676-023-09718-6

Matthias, A. (2015). Robot lies in health care: When is deception morally permissible? *Kennedy Institute of Ethics Journal*, *25*(2), 169–192. https://doi.org/10.1353/ken.2015.0007

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

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, *13*(6), 522–530. https://doi.org/10.1016/S2215-0366(25)00396-7

Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, *23*, 579–587. https://doi.org/10.1007/s10676-021-09596-w

Musiał, M. (2016). Magical thinking and empathy towards robots. In J. Seibt, M. Norskov, & S. S. Andersen (Eds.), *What social robots can and should do. Proceedings of Robophilosophy* (pp. 347–356). IOS Press. https://doi.org/10.3233/978-1-61499-708-5-347

Neely, E. L. (2014). Machines and the moral community. *Philosophy & Technology*, *27*, 97–111. https://doi.org/10.1007/s13347-013-0114-y

Nyholm, S. (2023). *This is technology ethics: An introduction*. Wiley Blackwell.

O'Neil, C. (2016). *Weapons of math destruction*: *How big data increases inequality and threatens democracy*. Crown.

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, *140*, Article 107600. https://doi.org/10.1016/j.chb.2022.107600

Prescott, T. J., & Robillard, J. M. (2020). Are friends electric? The benefits and risks of human-robot relationships. *iScience*, *24*(1), Article 101993. https://doi.org/10.1016/j.isci.2020.101993

Richardson, K. (2016). The asymmetrical "relationship." *ACM SIGCAS Computers and Society*, *45*(3), 290–293. https://doi.org/10.1145/2874239.2874281

Sætra, H. S., & Danaher, J. (2025). Resolving the battle of short- vs. long-term AI risks. *AI and Ethics*, *5*, 723–728. https://doi.org/10.1007/s43681-023-00336-y

Sandin, P., Peterson, M., Hansson, S. O., Rudén, Ch., & Juthe, A. (2002). Five charges against the precautionary principle. *Journal of Risk Research*, *5*(4), 287–299. https://doi.org/10.1080/13669870110073729

Scheutz, M. (2011). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, G. Bekey, & K. Abney (Eds.) *Robot ethics: The ethical and social implications of robotics* (pp. 205–221). The MIT Press.

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

Schwitzgebel, E. (2023). The Full Rights Dilemma for AI systems of debatable moral personhood. *ROBONOMICS: The Journal of the Automated Economy, 4*, Article 32.

Sebo, J. (2018). The moral problem of other minds. *The Harvard Review of Philosophy, 25*, 51–70. https://doi.org/10.5840/harvardreview20185913

Sebo, J., & Long, R. (2025). Moral consideration for AI systems by 2030. *AI and Ethics*, *5*, 591–606. https://doi.org/10.1007/s43681-023-00379-1

Shepherd, J. (2024). Sentience, Vulcans, and zombies: The value of phenomenal consciousness. *AI & Society*, *39*, 3005–3015. https://doi.org/10.1007/s00146-023-01835-6

Sinnott-Armstrong, W., & Conitzer, V. (2021). How much moral status could artificial intelligence ever achieve? In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking moral status* (pp. 269–289). Oxford University Press.

Smith, D. H., & Zeller, F. (2017). The death and lives of hitchBOT: The design and implementation of a hitchhiking robot. *Leonardo*, *50*(1), 77–78. https://doi.org/10.1162/LEON_a_01354

Sparrow, R. (2004). The Turing triage test. *Ethics and Information Technology*, *6*(4), 203–213. https://doi.org/10.1007/s10676-004-6491-2

Sparrow, R. (2017). Robots, rape, and representation. *International Journal of Social Robotics*, *9*, 465–477. https://doi.org/10.1007/s12369-017-0413-z

Strikwerda, L. (2017). Legal and moral implications of child sex robots. In J. Danaher & N. McArthur (Eds.), *Robot sex: Social and ethical implications* (pp. 133–151). The MIT Press. https://doi.org/10.7551/mitpress/9780262036689.003.0008

Stusińska, E. (2024). *Deus sex machina. Czy roboty nas pokochają?* WAB.

Sweeney, P. (2021). A fictional dualism model of social robots. *Ethics and Information Technology*, *23*, 465–472. https://doi.org/10.1007/s10676-021-09589-9

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research, 22*(3), Article e16235. https://doi.org/10.2196/16235

Turkle, S. (2010). In good company? On the threshold of robotic companions. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 3–10). John Benjamins. https://doi.org/10.1075/nlp.8.03tur

Van Gyseghem, A., Dierickx, K., & Barnhart, A. J. (2026). Consciousness and human brain organoids: A conceptual mapping of ethical and philosophical literature. *AJOB Neuroscience*, *17*(2), 78–92. https://doi.org/10.1080/21507740.2025.2519459

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & Society*, *36*, 487–497. https://doi.org/10.1007/s00146-021-01189-x

Warren, M. A. (1997). *Moral status: Obligations to persons and other living things*. Clarendon Press.

Weber-Guskar, E. (2021). How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners. *Ethics and Information Technology*, *23*, 601–610. https://doi.org/10.1007/s10676-021-09598-8

Weber-Guskar, E. (2022). Reflecting (on) Replica. Can we have a good affective relationship with a social chatbot?, In W. Loh & J. Loh (Ed.), *Social robotics and the good life. The normative side of forming emotional bonds with robots* (pp. 103–126). Transcript. https://doi.org/10.1515/9783839462652-005

Wei, M. (2025, November 27). The emerging problem of "AI Psychosis". Psychology Today. https://www.psychologytoday.com/us/blog/urban-survival/202507/the-emerging-problem-of-ai-psychosis

Weijers, D., & Munn, N. (2021). Human-AI friendship: Rejecting the appropriate sentimentality criterion. In V. C. Muller (Ed.), *Philosophy and theory of artificial intelligence* (pp. 209–223). Springer. https://doi.org/10.1007/978-3-031-09153-7_17

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*, W H Freeman & Co.

World Commission on the Ethics of Scientific Knowledge and Technology (COMEST). (2005). *The Precautionary Principle*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000139578

Żuradzki, T. (2021). Against the precautionary approach to moral status: The case of surrogates for living human brains. *The American Journal of Bioethics*, *21*(1), 53–56. https://doi.org/10.1080/15265161.2020.1845868