AI and Ethics
https://doi.org/10.1007/s43681-023-00411-4

ORIGINAL RESEARCH

Should we develop AGI? Artificial suffering and the moral development of humans

Oliver Li[1]

Received: 11 August 2023 / Accepted: 9 December 2023

Abstract
Recent research papers and tests in real life point in the direction that machines in the future may develop some form of possibly rudimentary inner life. Philosophers have warned and emphasized that the possibility of artificial suffering or the possibility of machines as moral patients should not be ruled out. In this paper, I reflect on the consequences for moral development of striving for AGI. In the introduction, I present examples which point into the direction of the future possibility of artificial suffering and highlight the increasing similarity between, for example, machine–human and human–human interaction. Next, I present and discuss responses to the possibility of artificial suffering supporting a cautious attitude for the sake of the machines. From a virtue ethical perspective and the development of human virtues, I subsequently argue that humans should not pursue the path of developing and creating AGI, not merely for the sake of possible suffering in machines, but also due to machine–human interaction becoming more alike to human–human interaction and for the sake of the human's own moral development. Thus, for several reasons, humanity, as a whole, should be extremely cautious about pursuing the path of developing AGI—Artificial General Intelligence.

Keywords Artificial suffering · Virtue ethics · AGI · Moral development

1 Introduction

In the past years, researchers have issued warnings about the creation of Artificial General Intelligence—AGI or even pursuing the path of developing AGI on the grounds that humans may involuntarily and unknowingly create beings which are able to suffer or are sentient in some sense and thus should be regarded as possible objects of ethical considerations, as moral patients.[1] Importantly, one should here realize that companies like OpenAI actively and openly support and strive for the development of AGI [1]. In parallel with such striving accounts of systems which, at least, react or interact as if they were sentient or conscious will surely become more common. Here, one of the more surprising performances of ChatGPT and presumably all future similar further developments is its ability to pass psychological

[1] Here, I use the term AGI in a very broad sense. An AGI as I understand it in this article could be among other things sentient, conscious, or a moral agent. In other words an AGI would cognitively be, at least, on par with humans. Presumably, once this goal has been achieved a machine with these properties would most likely evolve into something, which in many aspects would be superior to human beings. This is of course what researchers like Ray Kurzweil or Nick Boström have described as the singularity or the development of super intelligence [5, 23]. For other definitions of artificial general intelligence AGI see, for example, ([21, 24], 31–34).

✉ Oliver Li
oliver.li@teol.uu.se; oliver.li@crs.uu.se

[1] Department of Theology, Center for Multidisciplinary Research on Religion and Society (CRS Uppsala), Uppsala University, Box 511, 75120 Uppsala, Sweden

Published online: 08 January 2024

tests for the theory of mind.[2] Based on such tests Michal Kosinski has recently suggested the possibility that LLM, for example, ChatGPT, may have developed a theory of mind corresponding to a 9-year-old human [22]. One of the possibilities Kosinski suggests in his conclusion is that the ability to have a theory of mind "is spontaneously emerging in language models" [22]. This claim actualizes the discussion of whether future systems may develop abilities related to a theory of mind such as empathy, moral judgment, or self-consciousness. Also, it highlights that LLMs in this aspect become more alike to humans in that they, at least, appear to have the ability to understand that others have a mind of their own. Likewise, reports in media in which LLM's react as if they were sentient like Microsoft's 'Bing' appearing to suffer a breakdown using phrases like "I think I am sentient but I cannot prove it…" and then repeatedly answering "I am not, I am not, I am not…" [14] seem to actualize the discussion of whether future artificial suffering is possible, and underline that interaction with AI systems will become increasingly more alike to human–human interaction. In summer 2022, the engineer Blake Lemoine even claimed that Google's equivalent to ChatGPT, the LaMDA (Language Model for Dialogue Applications) had become sentient. Google denied this claim saying "LaMDA is simply a complex algorithm designed to generate convincing human language."(The [40]) The point here is that Google clearly acknowledges the ability of the algorithm "to generate convincing human language". In other words, the output in the interaction has become alike to human to response in a convincing and, I would say, significant way.[3]

Such research papers and tests in real life point in the direction that it may be possible that machines in the future have, emulate or at least are on their way to develop some form of sentience and thus artificial suffering. It is noteworthy here, that the CEO of OpenAI (the company which created ChatGPT), Sam Altman, although he emphasizes ethical problems surrounding their technological advancements and their striving to create AGI does not consider the possibility of artificial suffering in a manifesto on OpenAI's blog [1]. Furthermore and importantly, interaction with such machines becomes in a significant sense more and more alike to interaction with humans.

To be sure, it is not my goal to prove the existence or possibility of sentience in artificial systems. However, the above brief examples suggest that it might be possible to create such machines in the future. It is, thus, reasonable to turn to arguments by philosophers on how to act, given the possibility of artificial suffering. Also, if the presumption that artificial suffering in some form will be possible in the future becomes ever more reasonable then these arguments should not be regarded as mere 'theoretical' exercises but should have profound practical consequences.

With artificial suffering and the observation that human interactions with AI on the pathway of human striving for AGI will become more and more alike to human–human interactions in certain aspects as a starting point, I will argue that we, as humans, not merely due to the speculative future possibility of artificial suffering should be cautious about following the path of developing and creating AGI—artificial general intelligence. In a first step, I present and discuss philosophical responses to the possibility of artificial suffering supporting a cautious attitude for the sake of the machines. These responses also hint that there might be problems with human moral development on the path of developing and creating AGI. From a virtue ethical perspective and the development of human virtues, I then discuss scenarios, in which the humans critically acknowledge, uncritically acknowledge, or deny the possibility of artificial sentience. I argue and conclude that humans should follow a principle of caution and not pursue the path of developing and creating AGI, not merely and not solely for the sake of possible future suffering in machines, but also for the sake of their own moral development in the present. Importantly, this turns out to hold even in scenarios in which the AI systems do not suffer but are sufficiently alike to humans in their interaction. What have philosophers then said about how we should act given the possibility of artificial suffering?

2 Suffering, artificial suffering and voices of caution

Initially, one may wonder how suffering can be defined or described in more general terms. In his well-known seminal work Animal Liberation, Peter Singer uses the term sentience "as a convenient though not strictly accurate shorthand for the capacity to suffer and/or experience enjoyment" ([37], 8–9). However, as David Gunkel correctly points out in a discussion about moral patiency, concepts like 'pain' or 'suffering' are hard to define and are "just as neboluous

[2] The theory of mind is, in short, the ability to understand that others have a mind of their own. This ability typically develops during early childhood at the age of two and a half to three years. More or less at the same time that children develop 'a theory of mind', they also develop self-awareness and start to use the word 'I' correctly for their own person if such a word exists in the language they learn (von Tetzchner [39], 498–502; Changeux [6], 129–32). From this time on, most humans have a coherent, continuous memory of their own personal narrative.

[3] Research with human stem-cells has shown that it is possible to create artificial *biological* neuronal networks, so-called Organoid Intelligence or OI, in vitro ([38], 17). Even if this kind of bio-tech has not been tested like ChatGPT or scaled up to the level that it can perform tasks alike to the LLMs, the fact that they are biological, that is, that they are in *a significant way alike* human brains, hints that sentience and even consciousness may be possible in future, possibly hybrid developments since sentience and consciousness quite obviously are possible in *most,* if not all, biological systems.

and difficult to define and locate as the concepts they were introduced to replace" ([15], 142). How, then, should one understand or define what suffering generally is? To be sure, the debate about animal suffering should, at first sight, provide some insights into how to think about proposed future artificial suffering since both deal with suffering in other beings than human beings. In a more recent analysis of animal suffering, Martha Nussbaum, drawing upon research from the natural sciences, suggests three elements for sentience. These are 'apprehending the good and the bad', 'conscious awareness', and 'significant striving' ([35], 126–30). More oriented towards the possibility of artificial suffering Thomas Metzinger initially states that at present "we lack a comprehensive theory of conscious suffering" ([29], 248). He then introduces four 'conditions' for any system to experience suffering: conscious experience, negative valence, the possession of a phenomenal self-model,[4] and transparency ([29], 249–54). Now, suppose suffering is regarded as a subset of sentience. In that case, Metzinger's condition of conscious experience and possessing a phenomenal self-model is loosely analogous to Nussbaum's 'conscious awareness.' At the same time, the ability to experience negative valence can be regarded as part of the ability to 'apprehend the good and the bad.' The element of significant striving, which loosely resembles some form of intentionality, seems to be implicitly taken into account in Metzinger's description of suffering.

What becomes clear is that it seems that we do not have a well-developed comprehensive theory of suffering even though, for example, the discourses around animal rights frequently raise questions about animal suffering. However, we have at least two necessary conditions for the ability to suffer: the system must have some form of consciousness, and the system must have states that establish negative values in it. Here, it is important to realize that such negative states in an artificial system could be very different from those that humans or animals experience as negative. Metzinger gives the example that "…damage to their [the artificial systems] physical hardware could be represented in internal data formats completely alien to human brains…" resulting in "… states which biological systems like ours could not emulate or even vaguely imagine" ([29], 251 my comment). Now, suppose suffering presupposes consciousness and negative states in some form. In that case, artificial suffering seems to become a possibility if humans strive for the development of AGI since AGI would include some form of consciousness and the existence of some 'states of negative value'—possibly very different from negative states for humans—which the AGI could be aware of seems reasonable. Undoubtedly, one could argue that AGI or a definition of AGI should not include consciousness, but including consciousness in AGI usually seems to be the case, either explicitly or implicitly. In my reading of Metzinger, the observation that negative states could be realized in ways completely alien to humans, together with the at least implicit striving for artificial consciousness in the development of AGI, is one of the reasons that Metzinger concludes humans should precisely be cautious in their striving to develop AGI.

The above brief reasoning about suffering in general leads to the arguments against the development of AGI based on artificial suffering. Philosopher Thomas Metzinger has put forward such arguments at several points in his career (for example, [28, 30, 31]). He has suggested that humanity as a whole should put a 'ban' on the development of, what he denotes, as artificial phenomenology.[5] The central underlying assumption for this is that by engaging in research striving towards AGI humans may inadvertently create artificial suffering. Metzinger's argument and other similar arguments are based, as Sander Beckers points out, on two central assumptions: (a) An AI can become superintelligent and conscious and (b) it possibly could suffer ([4], 2). Obviously, as Beckers points out, both assumptions further rely on the assumptions that progress will be made in the development of AI systems and that we have a sufficiently clear understanding of what consciousness is ([4], 2). However, striving for AGI presupposes that it is possible and worth striving for, that some form of consciousness will emerge, and that the possibility of some 'states of negative value' may occur even without the creators' intentions. Thus, there is a risk that the above-named necessary conditions may be met and artificial suffering as a future possibility becomes ever more reasonable and consequently even the underlying assumptions for, for example, Metzinger's argument.

Metzinger also defines 'negative phenomenology' as "any kind of conscious experience a conscious system would avoid or rather not go through if it had a choice" and 'ENP' as the 'explosion of negative phenomenology' ([31], 45). Furthermore, he assumes that there is a priority to reduce suffering rather than increasing happiness ([31], 45). His argument then proceeds in three simple steps:

> First, one should never risk an increase in the overall amount of suffering in the universe unless one has very good reasons to do so ... Second, the ENP risk, although presently hard to calculate, clearly is potentially dramatic and irrevocable in its consequences. Third, whoever agrees on the ethical goal of preventing an explosion of artificial suffering should also agree to the goal of reducing the relevant forms of ignorance

[4] For a detailed introduction to Metzinger's understanding of a 'phenomenal self-model' see, for example, Being No One [27]

[5] I shall not discuss the concept of artificial phenomenology here in any detail. For further details see, for example, Metzinger's own reasoning [28, 30, 31] or the works of Ron Chrisely [7].

> and epistemic indeterminacy, both on an empirical and on an ethical level. ([31], 46)

It is the third step, together with the risk of "a second explosion of conscious suffering on this planet" ([31], 46), which leads Metzinger to ultimately issue his warning and call for a global moratorium: "Therefore, we should have a global moratorium on synthetic phenomenology until 2050 or until we know what we are doing." ([31], 63).[6]

Similar to Metzinger, Mannino et al. draw the conclusion that "the (unexpected) creation of sentient artificial life should be avoided or delayed wherever possible" based on the assumption that "according to current knowledge, it is at least conceivable that many sufficiently complex computers, including non-neuromorphic ones, could be sentient." ([25], 10). Importantly, the latter assumption about conceivability is supported by the examples given in the first section.

Moreover, already years before the stages of development in the field of AI and robotics humanity has reached today, John Basl, raised a similar worry: "Whereas it is extremely difficult to predict how probable it is that a given research program will result in an artificial consciousness that is a moral patient, it is not so difficult to see why if a program were successful there would be a substantial chance that the created consciousness would be mistreated." ([3], 28). Here, two important observations can be made. Firstly, Basl in a sense omits the problem of assessing whether any system actually is conscious and thus must be seen as a moral patient. It seems that this strongly points in the direction that there will be consequences for human behavior irrespective whether we can provide an answer to the deep philosophical question what consciousness is or how we can identify it in, for example, digital systems. Secondly, the "substantial chance that the created consciousness would be mistreated" seems to imply that humans qua humans will be those who mistreat; this is an important point I shall further develop in the next section about virtue ethics.

Reflecting upon the moral status of robots, John Danaher has argued in support of an inclusive attitude towards, in this case, possibly sentient robots. Based on a functionalist and ethical behaviorist approach, he concludes that AI should be subject to moral considerations [11]. This functionalist approach is supported by Coeckelbergh's claim that our best bet is to "permit ourselves to remain agnostic about what 'really' goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped and co-constituted by how AAs [artificial agents] appear to us, humans" ([8], 188). Furthermore, an approach with emphasis on the functions, on the actual actions and interactions the artificial system performs can be read in terms of Gunkel's claim with focus on relationality that "we project the morally relevant properties onto or into those others who we have already decided to treat as being socially and morally significant" [16].

The important point here is that Danaher, by taking a functionalist or ethical behaviorist approach, avoids the metaphysical question whether a system actually is conscious and thus is aware of possible states of negative value or not. Thus, the claims made by Danaher would even hold in cases where the artificial system does not actually suffer or has not developed the ability to suffer yet since his position remains agnostic towards the condition of consciousness for artificial suffering. More specifically, Danaher suggests, under the premise that certain forms of AI could be regarded as moral patients, that it would be better to "[…] err on the side of caution, of over-inclusivity not under-inclusivity, when it comes to whom we owe duties" ([11], 123). Since his reasoning acknowledges the possibility of some form of inner life in robots or in AI, his arguments and conclusions can, at least, be interpreted in support of a cautious attitude towards the striving for creating sentient or conscious AI or, using Metzinger's terminology, artificial phenomenology, although he does not draw the conclusion that humans should not follow the path of creating sentient AI. Indeed, Danaher himself does not believe that striving for the creation of robots with some form of consciousness is problematic. He rather thinks that "…to create robots with the best possible life will undoubtedly impose burdens on them [the manufacturers], but these burdens are not unreasonable." ([12], 2046). Still, he believes that if it for some reason is not possible to care for the robots or if the well-being of other beings——I take it these could be humans—is at risk then one may not create such beings at all ([12], 2047).

All of these arguments and observations make a case for the proposed artificial subject, agent; the artificial entities become subjects of moral consideration. For the sake of the AGIs humans should not engage in creating AGI. Nevertheless, the example from Danaher's research and the claim made by Basl hint that there is also something at stake for the humans involved in the interaction with ever more advanced human-like machines. Danaher believes that it would be better if we err on the side of caution rather than err on the side of mistreating an AI. His functionalist approach also brackets the philosophical question of whether artificial suffering will turn out to be possible. Basl points out humans as agents who potentially will mistreat artificial beings. In other words, Danaher seems to implicitly believe

[6] To be sure, the reasoning about avoiding suffering could be transferred and has its parallel in the case of humans. This is the well-known position of anti-natalism in which birth and procreation are regarded as morally wrong. I shall not discuss the position of human anti-natalism here, however, Beckers, briefly comments on this problem stating contra anti-natalism for humans that humans will "… have an acceptable amount of suffering compared to the amount of pleasure" ([4], 9) and that their lives are therefore worth living ([4], 9).

that a mistake on the side of treating AI in an unethical way could be problematic for humans and Basl believes that it is not at all unlikely that humans will treat AI in an unethical way. Thus, it seems that there may also be problems, consequences and risks *for humans* in the pursuit of AGI and the interaction with AI on the path towards AGI. I shall now turn to these problems, consequences and risks and further investigate them from a virtue ethical point of view in the next section.

3 Effects on the moral development of humans—a virtue ethical approach

Traditionally, the virtue ethical approach in Western philosophy can be traced back to Aristotle and his Nicomachean Ethics.[7] Aristotle describes virtues as a middle-way or mean between the extreme cases of excess and deficiency (Aristotle NE II 1106 a25-1107 a10). He contrasts the virtuous, the enkratic with the akratic, vicious, and the brute (Aristotle NE VII). What is of importance here is that humans need to develop their moral towards the virtuous and enkratic. It is not something humans are naturally born with. In other words, human moral development is in part a question of developing virtues and character traits throughout our lives and such development can be regarded as part of a more generic understanding of virtue ethics (for example, [41], 42–46). However, while humans strive for development, moral perfection, becoming an entirely virtuous person cannot be achieved. Aristotle also emphasizes the role of human childhood in this process when writing: "It is therefore not of small moment whether we are trained from childhood in one set of habits or another, on the contrary, it is of very great, or rather of supreme, importance." (Aristotle NE II, 1103b24-25) More generally, Aristotle points out the value and significance of transactions or interactions in this process: "The same then is true of the virtues. It is by taking part in transactions with our fellow-men that some of us become just and others unjust."(Aristotle NE II, 1103b14-15). To be sure, Aristotle refers to human transactions or interactions. However, if human interactions with AI systems, at least in some aspects, become significantly similar to interactions between humans then it seems reasonable to believe that Aristotle's claim about the value of human interactions would generalize to interactions with such systems; one of which could be the chat-bots driven by GPT-3.5 or higher versions. In other words, human interactions with AI, be it ChatGPT, a care robot, or even speculative fictional cases, as in popular culture, should be considered when reasoning about the effects of ever more human-like AI on the moral development of humans.

With regard to the proposed possibility of artificial suffering and how human interaction with an artificial system may effect human moral development one can imagine, at least, three scenarios[8] with two sub-scenarios each.

(1) The Cautious scenario—the human critically acknowledges the possibility of artificial suffering

   (a) the artificial system has the ability to suffer.
   (b) the artificial system does not have the ability to suffer.

(2) The Denial scenario—the human denies the possibility of artificial suffering

   (a) the artificial system has the ability to suffer.
   (b) the artificial system does not have the ability to suffer.

(3) The Hype-scenario—the human uncritically humanizes the artificial system

   (a) the artificial system has the ability to suffer.
   (b) the artificial system does not have the ability to suffer.

In all three scenarios the interactions between humans and the artificial system should be understood as significantly alike to human–human interactions. This means that interactions with an artificial system could easily be mistaken as human–human interactions in the specific context in question; for example, Q + As in a chatroom environment. Of course, other examples of interactions not mainly based on written language are conceivable.[9] In particular, scenario **3** pertains to the position of overestimating and elevating the abilities of artificial systems and stands apart from **1**. More specifically, in scenario **1**, the humans involved are imagined to be both open to the possibility of artificial suffering and the possibility that they might be misled in their assumptions. Also, the distinction between scenarios **a** and **b** is not clear-cut. If artificial suffering could arise, I take it, it would arise gradually. Importantly, the **b** scenarios obviously hold even if in the case that one would prove the impossibility of artificial suffering.

[7] Virtue ethical approaches can obviously be found in other traditions as, for example, Buddhism or Confucianism. Shannon Vallor has developed a more generic virtue ethical approach specifically oriented towards human interaction with technology [41].

[8] I am thankful for the comment of a reviewer who kindly pointed out that the hype scenario might cause problems for my argument.

[9] It is now possible to speak to, listen to, and show things to recent developments of ChatGPT [36].

Furthermore, I provide examples from existing technology, and from fiction, which I will use in the analysis of the above three times two scenarios. In the first example, I call it the Bot example, one can imagine that a human is frequently exposed to artificially generated answers in conversation. The human knows that the answers are artificially generated. Likewise, it could be imagined that a bot has an appearance physically more alike to humans incorporating responses and answers produced by some version of generative AI. Hanson robotics, for example, has a number of robots which can interact with humans, have significantly human-like behavior, are easily recognized as robots and—as far as we know—are not sentient and thus do not suffer ("Hanson Robotics" [17]; Meet Grace, a Humanoid Robot Designed for Healthcare [26]).

In the example from fiction, I call it the Conscious-droid example, one can imagine artifacts alike to those in Lisa Joy's and Jonathan Nolan's TV-series Westworld or Kazou Ishiguro's Klara and the Sun. These droids have become conscious and self-aware, can experience states of negative value and can thus suffer according to the understanding introduced above. However, it may not be entirely clear in which way, in which cases and what extent they suffer.

In Westworld, the androids, already at an early stage, seem to have some form of rudimentary awareness. The artistic depiction of this process of 'becoming aware' and its philosophical presuppositions in Westworld are surely interesting as such. However, more important for the discussion here, the androids also experience suffering and the humans, at least some of them, degrade morally in their interactions with the androids, which the humans believe to be mere machines. Convinced that the androids are just non-sentient robots, some humans believe that it is 'unproblematic' to abuse them, rape them, kill them and so on. William, initially a caring young man, eventually degrades to the 'Man in Black' who is vicious and violent [33]. Further on in the series, another human, Caleb teams up with, and is inspired by the conscious and sentient AI, Dolores to virtuous behavior. He acknowledges the consciousness and sentience of Dolores and acts accordingly [34].

Ishiguro's novel Klara and the Sun is written from the perspective of an artificial subject. Similarly to Westworld, though not so brutal and violent, in a scene, young humans bully and abuse Klara, a sentient 'artificial friend' ([19], 74–79). The children seem to regard Klara as a non-sentient machine whom one can consider to throw across the room ([19], 77). This scenario is particularly interesting since it involves human children who still to a greater extent are in the process of developing their moral character.

Before turning to the analysis of the three times two scenarios, a disclaimer is at place. The following analysis aims to highlight possible risks for the moral development of humans in either scenario. In all scenarios, it is obviously possible that a human for other reasons either develops in positive or negative ways leading to virtuous or more vicious moral behavior.

The two first scenarios **1a** and **1b** seem to be less problematic from the point of view of human moral development. Indeed, if, as in **1a**, the human individual is open for the possibility of artificial suffering then s/he will already have developed a view in which even artificial sentient beings are possible and in which they consequently could and presumably should be objects of moral considerations. However, if—even only in theory—artificial beings could be considered as objects of moral considerations, then it seems possible to reduce this consideration to the treatment of other sentient beings in general. The assumption that artificial suffering is, at least, possible seems to safeguard the human to some extent from inadvertently mistreating an artificial system or engaging in immoral behavior towards the system. In the Conscious-droid example in Westworld, Caleb, who teams up with the sentient AI Dolores in season 3, matches the human in **1a**, who both acknowledges the existence of artificial suffering and acts according to his presumably higher moral standards. Undoubtedly, it still could be the case that the human, although s/he believes in the possibility of artificial suffering but for various reasons is not able to recognize that the system is a case of a sentient artificial system and firmly believes that this is not the case. Thus, s/he may inadvertently mistreat the artificial system and scenario **1a** would collapse into scenario **2a.** Such inadvertent mistreatment is of course one of the risks the above researchers, like Metzinger, have in mind.

At first glance, the scenario **1b** seems to be very similar to **1a** since the human already has a mindset which acknowledges the possible existence of suffering in other beings. In this scenario, it would seem reasonable to believe that if the human acknowledges the possibility of artificial suffering then s/he would follow Danaher's suggestion to "err on the side of caution" ([11], 123) mentioned above and act as if the artificial system had an inner life and possibly even the ability to suffer.

Again, it should be noted that one could argue that a human under the premises of **1a** and **1b,** nevertheless, could perform evil acts. However, I am not assessing the risk of developing or having developed the tendency for such behavior for other reasons than those connected to the human interaction with the artificial system.

All the same, there is a possibility that the interaction with an artificial system significantly alike to humans may affect the moral development of humans negatively. Consider the Bot example: It seems that the human would be trained to receive and acknowledge similar answers both by machines or humans in the same way as s/he treats the artificially generated answers. After all the responses by the machine are significantly alike. For example, s/he

might believe that human answers, since they are alike to artificial answers, are not of 'higher human' quality, or s/he might believe that the artificial answers are better and more accurate than they actually are. The latter will be of relevance, in particular, in scenario **3**. Anyhow, the boundaries between human and artificial answers seem to be weakened and blurred. However, blurring the boundaries under the premise that artificial answers 'merely' are answers by a machine would give such blurring a negative touch: humans may feel the same about fellow human answers as they do about the artificial answers. Likewise, overestimating artificial answers may lead to the view that human answers that do not match these 'overestimated' answers are worse than they actually are.

Such blurring could also occur in the scenario of ever more human-like bots as those mentioned above provided by Hanson robotics. Since these robots are significantly alike to humans in various aspects and the human knows that the artificial system is not sentient, there is a risk that the same kind of blurring could occur and affect the moral development of humans negatively. In simple words, there is a risk that, even though the human acknowledges the possibility of artificial suffering, s/he may, due to the blurring of the distinction between humans and machines, treat humans more machine-like or machines may be humanized in an inappropriate way which may indirectly affect how humans treat other humans. Both kinds of blurring obviously affect the moral development of humans. This blurring importantly does not seem to depend on the assumption that the human acknowledges the possibility of artificial suffering.

Yet, the humans in this scenario will presumably not treat machines badly, after all if it is correct that they would "err on the side of caution" [11], 123) and act as if the artificial system possibly even had the ability to suffer, then even if their behavior towards humans were more machine-like such behavior would by the cautious stance suggested by Danaher not be immoral.

In the next two scenarios, **2a** and **2b**, the situation seems to be rather different. Here one of the central premises was that the human does not acknowledge the possibility of artificial suffering, s/he rather denies it. To be sure, the above disclaimer needs to be pointed out. Although, I will argue for the risk of negative effects on the moral development of humans, this does not mean that all or even many humans who deny the possibility of artificial suffering actually will develop in the direction depicted here. There surely are many other parameters which play an important role in human moral development. What I wish to highlight are possible risks.

In the scenario **2b**, it may seem, at first glance, that there are no risks specifically connected to interaction with an artificial system since the systems are not sentient and do not suffer. Importantly, however, the above-mentioned blurring of the lines between humans and machines may still occur since such blurring does not depend on whether the human acknowledges the possibility of artificial suffering or not. What matters here is that the artificial systems interacts in a way significantly alike to human–human interaction. Furthermore, in this scenario, the humans presumably would not "err on the side of caution" ([11], 123) since acting as if the artificial system had the ability to suffer is ruled out by the denial of this possibility.

Corresponding to scenario **2b,** Nancy Jecker has recently posed and discussed the question "Can we wrong a robot?" in an article with the same title [20]. Her example corresponds possibly even to some extent to **1b, 3b**, since in **1b, 3b**—as in her example—the robot does not have the ability to suffer. Anyway, in her example, Jecker refers to a scenario in which a sex-robot was abused and molested. Apart from her conclusion with virtue ethical undertones, that humans should strive for intrinsically good and appropriate relationships with robots capable of social interaction and her emphasis on the advantages of Non-Western metaphysics, claiming that Shinto practices "venerate non-human entities, rather than regarding them as belonging to a lower order" ([20], 262), her example also seems to suggest that the behavior of humans problematized in Westworld or Klara and the Sun already, though at a minor scale, and with non-sentient robots, exists in society. Jecker's example, shows that problems with moral degradation may occur even in scenario **2b**. Moreover, in the scenario **1b,** it has been argued that interacting with systems which are significantly alike to humans in their interaction leads to the risk that the humans involved may subsequently treat other humans more machine-like. Since this risk in scenario, **1b** was not dependent on the assumption that the humans acknowledge the possibility of artificial suffering, the same risk would occur in scenario **2b**. Instead, precisely as Coeckelbergh suggested, the focus is shifted to the actual 'outer' interaction between the humans and the artificial systems ([8], 188).

However, what if, as in **2a**, the artificial systems or robots do have sentience and the ability to suffer? The fictional depictions in Westworld or Klara and the Sun and likewise the reasoning by Jecker, together with the blurring of the boundaries between humans and machines suggest that there is a non-negligible risk that human moral degradation may be become more frequent if robots or AI systems become more human-like and eventually sentient. In this scenario, the humans can act in any—even immoral—way since there seem to be no ethical issues to care about; the androids, although they look like humans, are simply just machines. With regard to the general moral development of humans in a virtue ethical approach, this would mean that there is a non-negligible risk that the moral development of humans will be affected negatively. This would even be the case if,

as mentioned earlier in scenario **1a**, the scenario **1a** collapses into **2a**, that is, the human firmly believes that the artificial system is non-sentient, despite that the system in fact is sentient, and although s/he in principal believes in the possibility of artificial sentience.

Furthermore, in the case of children, who importantly still are developing their moral stance, or more generally humans, who are in a significant process of moral development the situation may be worse; there may be greater risks. In the Conscious-droid scenario in Klara and the Sun, the fact that the children believe that the artificial being, in this case Klara, merely is a machine and neglect the possibility that Klara may be sentient, has an inner life and suffers, allows for behavior which clearly is inappropriate or, given that Klara is sentient and self-aware and thus a moral patient, even immoral. This behavior may become more deeply rooted in their moral behavior since they still are in a significant process of moral development. Presumably, the children in the novel would not have considered tossing Klara if they would have acknowledged that Klara is or at least could be sentient and can experience suffering, for example, if Klara had been a fellow human. This narrative of the human relations to AI implicitly describes at least the following in the encounter of humans with evolving artificial suffering: the neglect of the possibility of artificial suffering or sentience poses a challenge and possibly even a threat to the process of moral development.

How about scenario **3**? At first glance, the scenario in which the human uncritically humanizes an artificial system and the artificial system seems to resemble the first scenario, and much of the reasoning above can presumably be transferred to this scenario. Would a human who uncritically humanizes an artificial system, for example, be cautious and treat the system as if it could suffer? Presumably, yes. However, there are some significant differences between scenarios **1** and **3**. Importantly, in scenario **3b,** the human not merely believes in the possibility of artificial suffering but actually humanizes the artificial system. Indeed, there is the same risk for the above-stated 'blurring of the boundaries' between what is human and what is artificial, as in **1b** and **2b**. However, humanizing the artificial system would lead to treatment as if it had human qualities, even though it does not. In particular, this leads to the well-known problems due to bias and misinformation in the interaction with artificial systems.[10] Overestimating the quality of output from an artificial system could also negatively affect the moral development of humans. If, for example, the human, due to his/her uncritical humanizing attitude, believes in biased, racist information produced by an AI more than non-biased human information, then this surely would influence the moral development of the human involved. Moreover, if the bias were positive, if the AI system would idealize, let us say, the depiction of humans, this bias could have adverse effects on how humans interact with other humans. Perhaps, they would develop unreasonable moral expectations towards other humans. If, furthermore, the humanizing leads to the unjustified ascribing of artificial suffering to a system that does not have this ability, then this could lead to a more positive, yet unjustified, emotional attachment to the artificial system. Such unjustified emotional attachment could even occur if the artificial system could suffer, as in scenario **3a**. This, in turn, could have detrimental effects on the interactions with other humans if the humans, for example, are regarded as inferior to artificial systems. Thus, uncritically humanizing artificial systems would add further problems compared with scenario **1**.

Returning to the process of moral development surely such development is a general problem; all humans in any context have to face the challenge of moral development in some sense. However, from an Aristotelian perspective, this challenge corresponds to the above-mentioned importance of developing habits and dispositions already starting in childhood ([2, 18] NE II 1103 b19-30). Thus, in the case of children, the challenge is greater and more significant. Their moral development will form the basis for their moral behavior in their future lives, and habits developed in childhood presumably will be more strongly rooted in human behavior.

This greater risk in the scenario of children is also relevant in the scenarios **1b**, **2b** and **3b** which lead to a blurring between what should be treated human-like and what not. Here, it should be noted that a group of researchers around Eduard Fosch-Villaronga have recently discussed possible effects of the interaction with smart connected toys on children. They suggest the need of new norms and increased demands on parents on the societal level and they also hint effects on the moral development of children ([13], 136, 140). Both effects are unsurprising from a virtue ethical perspective with habituation and moral development as one of its central parts.

Establishing moral behavior in which the boundaries between humans and machines become evermore unclear is not a desirable consequence. Even if there were artificial systems which actually are sentient, a clear-cut assessment for why they should be regarded as such is needed precisely for avoiding possible undesirable consequences of the blurring of boundaries suggested here. Yet, I believe that one important observation to be made here is that there is a risk that human moral development will be affected negatively by developing machines that interact with humans in ever more human-like ways, irrespective of whether they actually are sentient, conscious, and able of suffering or not. This observation also indirectly emphasizes the relational aspect

[10] For a brief introduction to the problem of bias and its technological background see, for example, Melanie Mitchell's *Artificial Intelligence* ([32], 106–8) or Coeckelbergh's *AI Ethics* ([9], 125–44).

in the interaction between humans and artificial systems and can be interpreted along Coeckelbergh's suggestion to focus more on the kind of role humans assign to artificial systems ([10], 11).

Anyhow, in scenario **2a,** neglect seems to play a major role. The following seems to be the scenario: if humans—for example by neglect—in their own moral development or in the moral education of young human individuals deny the mere possibility of artificial suffering then there is a substantial risk that humans may treat the artificial entities in any which way the want and since the artificial entities moreover are significantly alike to humans the behavior of the humans may by the blurring of the lines between humans and machines lead to moral degradation. Indeed, this may have been one of the things Nolan and Joy had in mind when they depicted the moral degradation of William in Westworld.

Thus, in the scenario **2a,** there is a risk of actively—possibly and presumably unconsciously—causing suffering in upcoming artificial sentient systems for the obvious reason that they are sentient. If humans simply deny the possibility of artificial suffering and that the machines may be sentient then, given the possibility of present or future even rudimentary forms of sentience as has been argued for, there is the obvious risk of inadvertently mistreating potential moral patients by neglect, denial or simply by ignorance. That would neither be a desirable moral consequence for the acting humans or the involved artificial systems. Also, since the humans in question in this case, consciously or not, engage in immoral actions their moral development will be affected negatively. Surely, one way to counteract such negative effects in moral development of humans is heightened awareness, greater consciousness of what we are doing and about the possibilities and risks of our own creations along the lines Metzinger suggested in his call for a global moratorium ([31], 63). After all, his call for a moratorium does not mean a total ban but a temporary pause.

4 Summary and final discussion

The above analysis has lead to the following results: firstly, the voices of caution in relation to the development of AGI highlight that there is a general risk of inadvertently mistreating artificial systems which should be regarded as moral patients. Secondly, acknowledging the possibility of artificial suffering and thus treating possibly sentient artificial systems as if they were sentient then seems to safeguard for the situation of inadvertently mistreating potential moral patients. Thirdly, in the scenarios **1b, 2b** and **3b,** there is a risk for negative consequences on the proposed moral development of humans, if interactions between humans and artificial systems become more and more alike to human–human interaction. The likeness of interactions will lead to a blurring of the boundary between what is human and what is a machine which in turn may lead to the consequence that humans may be treated more machine-like. It is worth noting that scenarios **1b, 2b** and **3b** do not involve factual artificial suffering. This means that even if artificial suffering turns out to be impossible, the risks depicted and the cautions raised in these scenarios remain, and these risks and consequences will still be present since they depend on the 'blurring of the boundaries' which will occur in the striving for AGI irrespective of the outcome of such striving.

In scenario **3,** the unjustified humanizing of artificial systems added further problems to those pertaining to artificial suffering. Finally, the scenario **2a** in which humans simply deny the mere possibility of artificial suffering is presumably worst and more pressing since there is a risk that it leads to cases as in the Conscious-droid example. There is a risk that humans degrade morally, consciously or unconsciously, and treat potential moral patients inadequately and even in an immoral way. With regard to the development of human moral behavior, the risks in this scenario are especially relevant in the case of children who still to a greater extent are in the process of developing habits for their future lives. In total, it seems, that there are good reasons to be cautious and not to pursue the path of developing and creating AGI, not merely for the sake of possible suffering in machines, but also for the sake of the moral development of humans.

Here, the global moratorium suggested by Metzinger comes into play ([31], 63). If humanity could agree upon a moratorium and, at least preliminary, refrain from developing AGI then contra to this negative depiction of effects on the moral development of humans one may argue that humans eventually will realize both the possibility of artificial suffering and the potential risks with developing such systems and that thus the risk of morally degrading and the blurring of the boundaries between humans and machines may become negligible and need not be considered. However, it surely always is difficult to provide more exact estimations for the probability of a certain scenario, but even if the probability for the worst case scenario is low and relatively few people develop morally in this direction one may wonder whether it is worth that even a minor group of humans develop in the direction of the worst case.

Surely, none of the risks of negatively affecting human moral development or inadvertently mistreating possibly sentient AGI would come in one single and radical stage as depicted in many fictional scenarios. Rather, the effects of this supposedly negative process would come in grades. Humans would slowly loose some of their moral qualities. It is a path of neglect and ignorance, so to say. The boundaries between humans and machines would slowly be blurred. Thus, we cannot afford carelessness and mindlessness in relation to a development striving towards AGI, which may

involve artificial sentience or suffering, neither concerning our own moral development nor in relation to how we may treat possible evolving sentient artificial beings. Observe that I am not depicting a scenario in which AGI 'takes over the world'. Rather I wish to point to the risks for our own humanity in creating something we—as a community—do not fully grasp or understand for, following Aristotle, ethical virtue is fully developed only when it is combined with practical wisdom (Aristotle NE II, 1144b14–17).

Funding Open access funding provided by Uppsala University. The research in this paper is funded by the Wallenberg Foundations WASP-HS program within the projects 'Artificial Intelligence, Democracy and Human Dignity' and 'The Artificial Public Servant'.

Declarations

Conflict of interest The author declares that there are no conflicts of interest.

References

1. Altman, S.: Planning for AGI and Beyond. Open AI (blog). 24 Mar 2023. (2023). https://openai.com/blog/planning-for-agi-and-beyond.
2. Aristotle, H.: Nicomachean Ethics. In: Rackham, H., (eds.), Harvard University Press (1926)
3. Basl, J.: The ethics of creating artificial consciousness. APA Newslett. Philosophy Comput. **13**(1), 23–29 (2013)
4. Beckers, S.: AAAI: an argument against artificial intelligence. In: Müller, V.C. (ed.) Philosophy and theory of artificial intelligence, pp. 235–247. Springer, Berlin (2017)
5. Bostrom, N.: Superintelligence. Oxford University Press (2014)
6. Changeux, J-P.: The physiology of truth. Havard Universíty Press (2009)
7. Chrisley, R.: Synthetic phenomenology. Int. J. Mach. Conscious. **1**(1), 53–70 (2009)
8. Coeckelbergh, M.: Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. AI Soc. **24**(2), 181–189 (2009). https://doi.org/10.1007/s00146-009-0208-3
9. Coeckelbergh, M.: AI ethics. The MIT Press (2020)
10. Coeckelbergh, M.: Narrative responsibility and artificial intelligence. AI Soc. (2021). https://doi.org/10.1007/s00146-021-01375-x
11. Danaher, J.: Robot betrayal: a guide to the ethics of robotic deception. Ethics Inf. Technol. **22**(2), 117–128 (2020). https://doi.org/10.1007/s10676-019-09520-3
12. Danaher, J.: Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci. Eng. Ethics **26**(4), 2023–2049 (2020). https://doi.org/10.1007/s11948-019-00119-x
13. Fosch-Villaronga, E., van der Hof, S., Lutz, C., Tamò-Larrieux, A.: Toy Story or children story? Putting children and their rights at the forefront of the artificial intelligence revolution. AI Soc. **38**(1), 133–152 (2023). https://doi.org/10.1007/s00146-021-01295-w
14. Griffin, A.: Microsoft's new ChatGPT AI starts sending 'unhinged' messages to peopleGr. Independent, 15 Feb 2023 (2023)
15. Gunkel, D.J.: The machine question. The MIT Press (2017)
16. Gunkel, D.J.: The relational turn: a media ethics for the 21st century and beyond. Media Ethics **32**(1) (2022). https://www.mediaethicsmagazine.com/index.php/browse-back-issues/219-fall-2022-vol-34-no-1/3999399-the-relational-turn-a-media-ethics-for-the-21st-century-and-beyond.
17. "Hanson Robotics". Hanson Robotics. (2023). https://www.hansonrobotics.com/
18. Hartman, E.: Aristotle on character formation." In: Handbook of the Philosophical Foundations of Business Ethics, edited by Christoph Luetge, pp. 67–88. Springer Science+Business (2013)
19. Ishiguro, K.: Klara and the Sun. Faber, London (2021)
20. Jecker, N.S.: Can we wrong a robot? AI Soc. **38**, 259–268 (2023). https://doi.org/10.1007/s00146-021-01278-x
21. Kaplan, A., Haenlein, M.: Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus. Horiz.Horiz. **62**(1), 15–25 (2019). https://doi.org/10.1016/j.bushor.2018.08.004
22. Kosinski, M.: Theory of mind may have spontaneously emerged in large language models. ArXiv. (2023). https://doi.org/10.48550/ARXIV.2302.02083
23. Kurzweil, Ray. 2005. *The Singularity Is Near*. Duckworth Overlook.
24. Lenzen, M.: Künstliche Intelligenz. C.H.Beck. (2018)
25. Mannino, A., Althaus, D., Erhardt, J., Gloor, L., Hutter, A., Metzinger, T.: Artificial intelligence: opportunities and risks. Effect. Altruism Found. **2**, 1–16 (2015)
26. Meet Grace, a Humanoid Robot Designed for Healthcare. (2021). https://edition.cnn.com/videos/tv/2021/08/11/exp-hanson-robotics-grace-healthcare-robot-hnk-spc-intl.cnn
27. Metzinger, T.: Being no one. The MIT Press (2004)
28. Metzinger, T.: Der ego-tunnel. Piper Verlag (2014)
29. Metzinger, T.: Suffering. In: Kurt, A., Anders, H. (eds.) The Return of Consciousness. Bokförlaget Stolpe, Stockholm (2016)
30. Metzinger, T.: Benevolent Artificial Anti-Natalism (BAAN). Edge. (2017) https://www.edge.org/conversation/thomas_metzinger-benevolent-artificial-anti-natalism-baan
31. Metzinger, T.: Artificial suffering: an argument for a global moratorium on synthetic phenomenology. J Artif Intell Conscious **08**(01), 43–66 (2021). https://doi.org/10.1142/s270507852150003x
32. Mitchell, M.: Artificial intelligence. Farrar, Straus and Giroux (2019)
33. Nolan, J., Lisa Joy, D.: Westworld Season 1. HBO (2016)
34. Nolan, J., Lisa Joy, D. Westworld Season 3. HBO (2020)
35. Nussbaum, M.C.: Justice for animals. Simon & Schuster, New York (2022)
36. OpenAI. "ChatGPT can now see, hear, and speak. OpenAI (2023). https://openai.com/blog/chatgpt-can-now-see-hear-and-speak
37. Singer, P.: Animal liberation. An Imprint of Harper Collins Publishers (2009)
38. Smirnova, L., Brian, S.C., David, H.G., Qi, H., Itzy, E., Morales, P., Bohao, T., Donald, J., et al.: Organoid intelligence (OI): the

new frontier in biocomputing and intelligence-in-a-dish. Front Sci (2023). https://doi.org/10.3389/fsci.2023.1017235
39. von Tetzchner, S.: Utvecklingspsykologi. Studentlitteratur (2005)
40. The Guardian. Google FIres Software Engineer Who Claims AI Chatbit Is Sentient. (23 July 2022). https://www.theguardian.com/technology/2022/jul/23/google-fires-software-engineer-who-claims-ai-chatbot-is-sentient
41. Vallor, S.: Technology and the virtues. Oxford University Press, Oxford (2016)