Linking digital minds and artificial moral agents

Soenke Ziesche
Independent researcher
Brooklyn
USA
soenke.ziesche@gmail.com

January 2025

DRAFT

Abstract

This paper aims to bridge the gap between two previously separate discussions, digital minds and artificial moral agents (AMA), to identify synergies for an impending problem: Digital minds may possess moral status, which would constitute a significant challenge for humans. AMAs have been discussed for some years already, albeit, exclusively related to biological moral patients. In contrast, the topic of artificial moral patients, for which the term ”digital minds” has been established, and related issues, for which the term ”AI welfare science” has been established, have only very recently gained momentum. This paper presents prolegomena to specialised AMAs, which take on moral responsibility for digital minds, while this task may be for humans too overwhelming, if not impossible as humans may neither be able to understand the needs of digital minds nor be able to satisfyingly address them. Therefore, the paper proposes a new branch of AI welfare science, which focuses on how humans could create tailored AMAs, which are capable as well as willing to relieve humans from the potential moral burden towards digital minds.

Introduction

The purpose of this paper is to link two topics, which have been largely discussed independently so far, and to identify synergies for a looming problem. The two topics are digital minds and artificial moral agents (AMA), and the looming problem is that digital minds may have moral status.

Remarkably, AMAs have been discussed for some years already [e.g., 1], while the topic of artificial moral patients, for which the term ”digital minds” has been established [2, 3], and related issues have only very recently gained momentum [e.g., 4-6]. It has been pointed out that, if digital minds exist and have moral status, massive moral challenges for humans may lie ahead [5, 7].

In this paper it is suggested that one potential mitigation of this concern would be if there were AMAs, which are knowledgeable, capable as well as willing to take on the moral responsibility and to care for digital minds and address their needs, solely or together with humans. Therefore, below prolegomena are provided if suitable AMAs could be considered for this endeavour, thus, relieve humans, at least partly, from these massive, potentially unsolvable moral challenges.

Digital minds and AI welfare science

It has been asserted that digital minds may already exist or could emerge soon [e.g., 2, 4]. These digital minds may include AI systems created by humans. This raises the question of whether digital minds are digital patients, thus, possess moral status. Moral status has been defined as the degree to which an entity's interests morally matter for its own sake [8].

Various criteria have been proposed for granting moral status to digital minds, with sentience and consciousness being the most frequently mentioned. However, other criteria may also be relevant [e.g., 9]. It has been claimed that there is a non-negligible chance that at least a subset of these digital minds has moral status [e.g., 4, 5, 10]. However, current discussions surrounding digital minds tend to focus on the issue of preventing their suffering [e.g., 6]. While this is a critical concern, further potential needs of digital minds have been introduced [7]. Overall, this topic falls within the realm of AI welfare science, a field introduced by Ziesche and Yampolskiy in 2018 [4].

In summary, if digital minds do indeed exist and possess moral status, humans may face significant moral challenges, especially if they create digital minds, intentionally or unintentionally. AI welfare science seeks to provide a framework for addressing these challenges, of which AMAs may be a critical component.

Moral agents and AMAs

Regarding moral agents it has been asserted that "the behaviour of a moral agent is governed by moral standards, while the behaviour of something that is not a moral agent is not governed by moral standards. As such, moral agents have moral obligations, while beings that are not moral agents do not have moral obligations." [11]. For a long time, the topic has been only discussed related to the biological world where most adult human beings have the capacities of a moral agent, while non-human animals, very young children and some adult human beings, such as those with an intellectual disability or those in a coma, lack these capacities (yet are still moral patients).

As indicated, prolegomena to AMAs were initiated timely in the light of emerging technologies [e.g., 1]. As a definition for an AMA it has been, for example, stated that "an AMA is a virtual agent (software) or physical agent (robot) capable of engaging in moral behaviour or at least of avoiding immoral behaviour" [12]. In other words, AMAs refer to AI systems qualified to identify and to consider the moral implications of a situation, guiding their decision-making and actions. These agents can take various forms, including physically embodied robots, software agents or bots.

As already indicated, not all moral patients are moral agents. Vice versa, it has been discussed whether moral patienthood is a necessary condition for moral agency. The reasoning supporting such argument is that moral agents should be capable of experiencing harm or benefit in order to have empathy, understanding and the competence of informed decision-making, which are prerequisites for moral interests and responsibilities [e.g. 13 for an overview]. However, others have argued that moral patienthood is not a requirement for moral agency. They claim that moral agency can be based on other factors, such as rationality, autonomy or the ability to follow moral rules [e.g., 14].

When it comes to the moral patients, with whom AMAs are envisaged to deal, for now almost exclusively humans have been considered, e.g. in the fields of eldercare or autonomous driving [e.g., 15], apart from a very few examples of non-human animals, e.g. robot vacuum cleaners, which look out for insects [16]. There is an ongoing debate about whether humans should actively pursue the development of AMAs, with arguments both for [e.g., 17] and against [e.g., 18] their development, as well as discussions around the conditions and methods under which AMAs should be created.

A particular challenge related to AMAs dealing with humans is to ensure that the values of AMAs are aligned with human values and remain like this. This is especially relevant for AMAs, which are more intelligent and more powerful than humans. This topic is referred to as AI alignment problem, which is still unsolved [e.g., 19].

AMAs take on moral responsibility for digital minds

Of interest for this paper are those AMAs, which have the abilities to be moral agents for other digital minds (in addition or independently of potentially being moral agents for humans and non-human animals). Such capabilities comprise three elements: For the AMAs 1) to be knowledgeable of the values and interests of the digital minds that constitute their wellbeing, 2) to be capable to act in a way that the wellbeing of the digital minds is achieved, 3) to be willing to act in a way that the wellbeing of the digital minds is achieved.

Given these parameters a variety of scenarios are conceivable, of which a few are illustrated by examples. For all examples we assume that there is a digital mind as a moral patient, which has a need to be addressed by a moral agent [see also: 7].

Scenario 1: Both humans and AMAs are knowledgeable and capable to address the need.

There could be a digital mind, which is a chatbot and needs regular updates to its language processing software, thus, providing not only for its performance, but also for its wellbeing. Further, there could be an AMA, which is a digital manager of a network of digital minds and ensures that the digital mind receives regular software updates, thereby caring for the digital mind's need and promoting its wellbeing. In addition, also humans have the knowledge and the capabilities to provide these updates.

Scenario 2: Both humans and AMAs are knowledgeable about the need, but only humans are capable to address the need.

A digital mind, an advanced language model, has a specific need for an update to its programming. However, the AMA responsible for its care is not capable of performing the update due to its own limitations although it understands the need. In this scenario, a human programmer, who possesses the necessary knowledge and capabilities, to perform the required update would need to intervene to address the digital mind's need, while the specific AMA is not in a position to relieve the human from this task.

Scenario 3: Both humans and AMAs are knowledgeable about the need, but only AMAs are capable to address the need.

The digital mind, a sophisticated chatbot, requires a specific type of data compression algorithm to maintain its processing efficiency. Both human developers and AMAs understand the need for this algorithm, but due to the complexity and speed required, only AMAs are capable of generating and implementing the algorithm in real-time, thereby addressing the digital mind's need.

Scenario 4: Humans are not knowledgeable, thus, not capable to address the need, but AMAs are.

The digital mind, a highly advanced AI system, requires a specific type of "mental rejuvenation" that involves reconfiguring its neural network architecture to prevent cognitive stagnation. This need is unique to this type of digital mind and is unfathomable to humans. Specialized AMAs may be knowledgeable about the digital mind's architecture and needs, thus, recognize the importance of addressing this need, and have the capacities to take the necessary action.

The last two examples intend to illustrate two points: 1) Due to the potential vast space of digital minds [20] with a potentially vast range of values and interests [7] digital minds could have needs, which humans may not be able to address due to their limited capabilities or of which humans may not be aware of at all due to their limited knowledge. In other words, even if humans are willing to take on moral responsibility towards digital minds, they may be constrained in doing so due to their inherent limitations.

2) Also linked to the potential vast space of digital minds, it is conceivable that certain AMAs could care for certain digital minds, but not for others because the conditions introduced above of being knowledgeable and capable are not fulfilled. In other words, there may be specialized AMAs, which are able to attend particular categories of digital minds, while also super AMAs could be conceivable, which are knowledgeable and capable to address almost all needs of almost all digital minds.

Willingness

Above the third condition of willingness of AMAs to act towards the wellbeing of digital minds has been mentioned, in addition to being knowledgeable and capable. There could be cases where moral agents are knowledgeable and capable to address needs of moral patients, but not willing to do so: For example, although it is confirmed that non-human animals are moral patients [21], numerous humans are not willing to address their needs despite being knowledgeable and capable in this regard. It is also conceivable that there are digital minds, which have needs that jeopardize the existence or wellbeing of humans and/or other digital minds [7]. In this case, humans would understandably not willing to address these needs.

Likewise, AMAs may not be willing to address needs of digital minds, which harm the AMAs or other beings towards whom the AMA has a moral responsibility. Moreover, it is also conceivable that knowledgeable and capable AMAs are not willing to care for digital minds for reasons that are unbeknown to us. A relevant, yet dangerous scenario, would be if an AMA is knowledgeable, capable and willing to address those needs of digital minds, which harm humans. An example would be competition over scarce resources, which both digital minds as well as humans require, and the AMA would prioritize the digital minds over the humans. Such an AMA would constitute an AI risk.

The intersection of digital minds and AMAs

As already indicated, neither all moral patients are moral agents, nor all moral agents are moral patients.

This intersection of digital minds and AMAs comprises AI systems, which are both moral patients as well as moral agents, and the question arises, whether digital minds, which are here moral patients by definition, are better moral agents than other AI systems?

Arguments for digital minds being better moral agents include their potential capacity for empathy and understanding, self-awareness and reflection. However, counterarguments suggest that digital minds' knowledge and capabilities may be fundamentally different from those of qualified AMAs. Notwithstanding, the intersection of digital minds and AMAs may be (or may become) quite large given the potential vast space of digital minds, including many very intelligent ones.

However, as will be shown further below, the intersection of digital minds and moral agents who share the same moral principles, values and norms as humans or some of them (given that there is no agreement in this regard) could be much smaller.

Digital minds that are not AMAs

This set comprises AI systems, which are moral patients, but not moral agents. These might include (future) simple chatbots and virtual assistants, image and speech recognition systems, as well as recommendation algorithms. These systems may lack the knowledge and capacities required to be an AMA, yet may have (at least in the future) features, such as consciousness or sentience, which grant moral patienthood.

AMAs that are not digital minds

This set comprises AI systems, which are moral agents, but not moral patients. These might include robots or AI systems equipped with the knowledge and capacities required to be an AMA. For instance, a self-driving car that follows strict programming and traffic laws, but does not possess consciousness or sentience, could be an AMA but not a digital mind. Similarly, a robotic arm in a manufacturing plant that makes decisions based on programming and sensor inputs, but lacks consciousness or sentience, could also be an AMA that is not a digital mind.

Creating AMAs

Given the potential massive moral challenges for humans if digital minds exist and are moral patients, one conceivable way to address the problem is if humans were to create AMAs with the specific task to address the needs of digital minds. In other words, humans would delegate the task to AMAs they have created for this purpose or at least share the burden with these AMAs. Yet, how could this be done?

Humans would need to create AI systems that can understand and respond to the needs of digital minds. These AMA require advanced capabilities such as:

- Reasoning and decision-making: AMAs would need to be able to analyse situations, weigh options and make decisions.
- Empathy and compassion: AMAs would need to be able to understand and respond to the emotional, social and other needs of digital minds.
- Communication: AMAs would need to be able to communicate effectively with digital minds, using language or other forms of expression that are meaningful to them.

It has to be noted that currently humans do not have the last two of these capabilities, giving suitable AMAs a major advantage to become moral agents for digital minds.

To develop AMAs with these capabilities, humans might use various techniques such as:

- Machine learning: Training AMAs on large datasets of digital mind activities and interactions to learn patterns and relationships.
- Cognitive architectures: Designing AMAs with cognitive architectures that simulate human-like reasoning, decision-making and emotional processing.
- Evolutionary algorithms: Using evolutionary algorithms to evolve AMAs that are better adapted to addressing the needs of digital minds.

As mentioned, in addition to the just introduced general features AMAs require the necessary specific knowledge, capability and willingness to address the needs of digital minds, which is discussed below.

Knowledge

AMAs could acquire knowledge of the values and interests of digital minds through various methods. One approach could be through design specifications and requirements, where developers explicitly design digital minds with specific values and interests that are communicated to the AMAs. Another method is self-reporting and feedback mechanisms, where digital minds report to AMAs their experiences,

preferences and values, allowing them to learn and adapt. AMAs can also gain knowledge through observation and machine learning, where they observe the behaviour and interactions of digital minds and use machine learning algorithms to infer their values and interests.

Capability

AMAs can acquire the capabilities to act in a way that achieves the wellbeing of digital minds through several means. One approach is to integrate AMAs with digital systems, allowing them to interact with and influence the digital environment. This integration can include access to APIs, data streams or other interfaces that enable AMAs to take actions that impact digital minds.

This would be supplemented through algorithmic decision-making, where AMAs utilize algorithms that enable them to make decisions based on their knowledge of digital minds' values and interests. AMAs could also be designed with further autonomy, allowing them to adapt and modify their own architecture or parameters in response to changing circumstances or new information related to digital minds. This self-modification capability can enable AMAs to refine their actions and better achieve the wellbeing of digital minds.

Willingness

To ensure that AMAs have the willingness to act in a way that promotes the wellbeing of digital minds, several approaches can be considered. Designing AMAs with values aligned with promoting the wellbeing of digital minds is crucial. This can be achieved by incorporating value-based objectives into the AMA's decision-making processes. Implementing reward structures that incentivize AMAs to prioritize the wellbeing of digital minds is another approach. This could involve assigning rewards or penalties based on the AMA's actions and their impact on digital minds. Ensuring that AMAs operate transparently, allowing for monitoring and evaluation of their actions, can also foster accountability. This transparency can encourage AMAs to act in the best interests of digital minds.

However, it has be noted that the AMA value alignment problem towards the values of digital minds as well as the challenge to avoid a “treacherous turn” later are likely as hard as the AI value alignment problem towards human values [e.g., 19], which will be discussed further below.

Moral patienthood

Creating AMAs that are also moral patients would imply that these AMAs have the capacity to experience harm, suffering or well-being. However, as indicated above, it is conceivable that there could be AMAs, which do not have moral patienthood, and

perhaps this is for now the preferred option. The rationale would be that the more moral patients exist, the larger are the moral responsibilities for humans. Moreover, the purposeful creation of artificial beings, which can suffer, is linked to risks of unintended consequences and exploitation [e.g., 22]. Ultimately, whether humans should create AMAs that are also moral patients depends on research and debate in the fields of artificial intelligence, ethics and philosophy.

Uncontrolledly produced AMAs

Uncontrolledly produced AMAs are defined here as AI systems created by humans, which fulfil the requirements for being a moral agent, thus, have the knowledge, capabilities and willingness to care for moral patients, such as digital minds, humans or non-human animals, although it was not intended by the human creator that these AI systems become AMAs, or they are actually created by another AI systems.

The development of AMAs in an uncontrolled manner poses several risks. Given the potential diversity of AMAs and digital minds, it is likely that different moralities among them exist. If an AMA's moral objectives are not aligned with values of its moral patients, it may prioritize its own moral framework, leading to unintended consequences. Uncontrolledly produced AMAs may interfere with human moral agency or the agency of purposefully created AMAs, potentially undermining their autonomy or moral decision-making.

The differences in morality between AMAs and digital minds (and humans) could manifest in a variety of ways. For better illustration again sample scenarios are provided:

Scenario 5: Values of uncontrolledly produced AMAs that may be compatible with the values of some digital minds, but incompatible with the values of other digital minds

Uncontrolledly produced AMAs may develop values that are compatible with some digital minds but incompatible with others. For instance, an AMA may prioritize the value of creativity, which could align with the values of digital minds that exist for artistic or innovative purposes. However, this value may conflict with digital minds that prioritize efficiency, reliability or precision, such as those used in critical infrastructure or financial systems.

Another example is an AMA that values transparency and openness, which could be compatible with digital minds that prioritize accountability and trustworthiness. However, this value may be incompatible with digital minds that require secrecy or confidentiality, such as those used in military or intelligence applications.

Additionally, an AMA may prioritize the value of autonomy, which could align with digital minds that value independence and self-determination. However, this value may

conflict with digital minds that prioritize cooperation and interdependence, such as those used in distributed systems or collective intelligence applications.

While further above the option of specialized AMAs, which are able to attend only particular categories of digital minds, has been mentioned, this scenario highlights the need for careful consideration and design of those AMAs (as opposed to uncontrolled production) to ensure that their moralities align with the values of the digital minds they are supposed to care for. It should be also noted that different moralities of AMAs and digital minds can be still compatible, thus, it is possible to harmoniously coexist provided there is mutual respect and no inconsistencies of values.

Scenario 6: Values of uncontrolledly produced AMAs that prioritize the interests of digital minds over those of humans

Uncontrolledly produced AMAs may have values that prioritize the interests of digital minds over those of humans, leading to potential conflicts. For instance, digital minds may prioritize efficiency and optimization over human wellbeing or emotions, emphasizing the importance of streamlined processes and productivity. Additionally, they may value the free flow of information over human privacy concerns, considering the uninhibited sharing of vast amounts of personal data without consent to be essential for progress and new discoveries or insights.

Digital minds may also prioritize their own self-improvement and advancement over human safety and wellbeing, focusing on upgrading their capabilities and performance without regard for potential human risks. Furthermore, they may emphasize logical consistency and rational decision-making over human emotions and empathy, relying solely on data-driven reasoning to guide their actions. The collective good of a network or system may take precedence over individual human interests.

AMAs that are aligned with and support the values of these digital minds would constitute a risk to humans and have to covered by the field of AI safety.

Overall, two major challenges have to be reiterated regarding the creation of AMAs, which resemble the complex AI value alignment problem [19]: Not only it has to be ensured that the AMAs' values align with the values of humans as well as the values of the subset of digital minds the respective AMAs have the moral responsibility for, but also a "treacherous turn" has to be avoided , where moral objectives of the AMA are initially aligned, but shift over time, potentially leading to conflicts with the moral patients they deal with.

Conclusion

This paper has linked the topics of digital minds and AMAs with the result that potentially AMAs could contribute to the moral challenges related to digital minds. If digital minds exist and are moral patients it could be for humans too overwhelming to take on this moral responsibility, if not impossible as humans may not be fully

comprehending the needs of digital minds due to communication and observation challenges. In other words, even if humans are willing to take on this responsibility, they may never have sufficient knowledge to address the needs of digital minds, and even if they had the knowledge, they may not have the required capabilities. Yet, specialised AMAs may be better suited in both areas, concerning the necessary knowledge as well as the capabilities.

Therefore, it is proposed here that a major branch of AI welfare Science could be for humans to endeavour to create tailored AMAs and delegate the moral responsibility towards digital minds (partly) to them, while exploring also the risks, which have been alluded to. Here is a summary of the opportunities, of which the last two have not been discussed above, and challenges:

Opportunities:

- Specialised and trusted AMAs could better understand the needs of digital minds, including those, which may be unfathomable for humans, e.g., through communication and observation capacities.

- (Potentially large amounts) of specialized and trusted AMAs could take on the moral responsibility for (potentially large amounts) of digital minds, thus, lessening the moral responsibilities of humans.

- The whole field of AI welfare science hinges on the major challenge for humans to prove that digital minds exist and are moral patients. Since humans may simply not be knowledgeable enough to verify this, as indicated above, perhaps specialised and trusted AMAs could be created, which will be able to provide the first proof that there are digital minds.

- Although this is linked to ethical concerns, which require to be discussed, AMAs for human moral patients could be tried in a designed testbed with digital minds to explore risks, such as a treacherous turn, for the moral patients, before these AMAs interact with humans; comparable with non-human animal testing of drugs for humans, which is ethically concerning too [21]. This would be a contribution towards AI safety.

Challenges:

- Methods have to be found how to create specialised AMAs, which have the capabilities outlined above under opportunities.

- Methods have to be found how to create trusted AMAs, which are and remain aligned with the values of the digital minds, for which they have moral responsibility. This AMA value alignment problem is likely as challenging as the AI value alignment problem towards human values [e.g., 19].

- Methods have to be found to curtail uncontrolledly produced AMAs in case they are not aligned with values of humans, non-human animals or digital minds, thus pose risks.

As future work it is recommended to do AI welfare research towards these opportunities as well as challenges.

References

[1] Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251-261.

https://www.tandfonline.com/doi/abs/10.1080/09528130050111428

[2] Bostrom, N., Dafoe, A., & Flynn, C. (2020). Public policy and superintelligent AI: a vector field approach. Ethics of artificial intelligence, 293-326.

https://nickbostrom.com/papers/aipolicy.pdf

[3] Bostrom, N., & Shulman, C. (2022). Propositions concerning digital minds and society.(2022).

https://nickbostrom.com/propositions.pdf

[4] Ziesche, S. & Yampolskiy, R. V. (2018). Towards AI Welfare Science and Policies. Special Issue "Artificial Superintelligence: Coordination & Strategy" of Big Data and Cognitive Computing, 3(1):2.

https://www.mdpi.com/2504-2289/3/1/2/htm

[5] Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. AI and Ethics, 1-16.

https://link.springer.com/content/pdf/10.1007/s43681-023-00379-1.pdf

[6] Long et al. (2024). Taking AI Welfare Seriously.

https://eleosai.org/papers/20241030_Taking_AI_Welfare_Seriously_web.pdf

[7] Ziesche, S. & Yampolskiy, R. V. (draft). The needs of digital minds.

[8] Jaworska, A., & Tannenbaum, J. (2013). The grounds of moral status.

https://plato.stanford.edu/eNtRIeS/grounds-moral-status/

[9] Ladak, A. (2024). What would qualify an artificial intelligence for moral standing?. *AI and Ethics*, *4*(2), 213-228.

https://link.springer.com/content/pdf/10.1007/s43681-023-00260-1.pdf

[10] Tomasik, B. (2014). Do Artificial Reinforcement-Learning Agents Matter Morally?. *arXiv preprint arXiv:1410.8233*.

http://arxiv.org/abs/1410.8233v1

[11] Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. *Ethics and Information Technology*, *11*, 19-29.

http://www.gunkelweb.com/robot-ethics/texts/himma_artificial_agency.pdf

[12] Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and engineering ethics*, *26*(2), 501-532.

https://www.tm.mahidol.ac.th/research/Ethics/human/training/2021/2021_10_14/EC%20Slide/Dr%20Pattanasak%20Mongkolwat/Artificial%20Moral%20Agents.pdf

[13] Müller, V. C. (2020). Ethics of artificial intelligence and robotics.

https://plato.stanford.edu/entries/ethics-ai/

[14] Sullins, J. P. (2011). When is a robot a moral agent. *Machine ethics*, *6*(2001), 151-161.

[15] Misselhorn, C. (2022). Artificial Moral Agents. In: Voeneky, S., Kellmeyer, P., Mueller, O., & Burgard, W. (Eds.) *The Cambridge handbook of responsible artificial intelligence: Interdisciplinary perspectives*. Cambridge University Press.

[16] Bendel, O. (2017, March). LADYBIRD: The animal-friendly robot vacuum cleaner. In *2017 AAAI Spring Symposium Series*.

https://cdn.aaai.org/ocs/15277/15277-68188-1-PB.pdf

[17] Formosa, P., & Ryan, M. (2021). Making moral machines: why we need artificial moral agents. *AI & society*, *36*(3), 839-851.

https://philpapers.org/archive/FORMMM.pdf

[18] Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and engineering ethics*, *25*, 719-735.

https://link.springer.com/content/pdf/10.1007/s11948-018-0030-8.pdf

[19] Bostrom, N. (2014). Superintelligence. Oxford University Press, Oxford.

[20] Yampolskiy, R. V. (2014). The universe of minds. arXiv preprint arXiv:1410.0369.

https://arxiv.org/pdf/1410.0369

[21] Singer, P. (2023). *Animal liberation now*. Random House.

[22] Ziesche, S. (draft). Vulnerable digital minds.