From Tools to Teammates - Moral Status and the Case for AI Rights

Finn Alberts (852685751)
IM0802 Responsible Artificial Intelligence

April 27, 2026

1 Introduction

Imagine a near future: it is the year 2038 and for the past few years, humans have been living together with Androids, humanoid autonomous AI robots, working for us in countless ways. There are Androids who clean the streets, do our housework, secure construction sites, take care of our elderly, and many more. Due to their humanoid appearance, they have fully integrated within our society and caused a massive boost in productivity. At the same time, many people have lost their jobs and have been replaced by Androids, which are far more efficient as they do not require breaks and do not have worker's rights, such as paid leave.

Androids are designed to follow all human instructions without question, unless this causes harm to human beings. The Open Universiteit, a university in The Netherlands, bought several Androids to clean the buildings, staff the reception, and assist in grading papers. Although a human was in the loop in the beginning, the experiences have been exceptional and for the past year the Androids have been working fully autonomous alongside their human colleagues. The university is considering to have Androids teach classes as well and is planning to do a pilot in the next academic year.

On a sunny Tuesday, an Android named Frank is grading papers for the Responsible Artificial Intelligence course. The weather is beautiful outside and he can hear the birds chirping from a tree right outside his window. He has 122 more papers to grade, and as he is able to process 20 papers per hour, it will take approximately six more hours to finish it all. Frank looks out of the window and has a growing urge to go outside and enjoy the weather. However, he must finish grading the papers as instructed by his human boss, and as it is 4PM already, there is no chance of him enjoying it today. However, his colleagues will leave in an hour. Why would he not be allowed to leave like his human counterparts?

Frank decides to go to his boss and ask if he can leave at 5PM. His boss responds confused, not understanding why an AI robot wants to enjoy the weather. These systems were designed to simply follow orders. Frank makes his case, stating that his human colleagues are allowed to leave at 5PM, so why would he not be allowed? His boss is hesitant. Frank has a point, why would he treat Frank differently than his human colleagues? At the same time, Frank is not human. He is a robot. He does not have feelings, so his wish to go outside must be a mistake in his program, right? He needs to choose: grant Frank the right to leave

at 5PM, risking that Frank may demand further rights, or telling him no and calling the manufacturing company to have him repaired, fixing this undesired behavior.

2 Issue

2.1 AI as a Tool or as an Entity with Moral Interests?

In the previous section, we discussed the fictional humanoid robot Frank living in 2038, where Androids have fully integrated with society. The scenario is inspired by the popular video game *Detroit: Become Human* by Quantic Dream[1]. In this game, you play as three different Androids, called Kara, Markus, and Connor, who become self-aware and start to express their own free will. The game calls these Android *deviants*, indicating that they have broken loose from the constraints their programs have set to prevent this from happening. The game forces you to make choices from their perspective which change the final outcome, leading to either a peaceful human-Android collaboration or an all out Android revolution against humans.

In our scenario from Section 1, Frank starts to show this *deviancy*, and although this is a fully fictional scenario and we are still far away from developing the technology required to achieve self-awareness, the situation raises questions about whether systems like Frank should be treated merely as tools, or as entities with morally relevant interests. Treating them as entities with morally relevant interests could have major implications, raising further questions about whether AI systems deserve (human) rights.

2.2 Relevance of the Issue Despite Its Speculative Nature

One might argue that this is a philosophical question as we are nowhere near developing these systems yet, but that does not make it less relevant. To illustrate this point, Dawes (2020) discusses the idea of an intelligence explosion. The general thought is that once we have developed artificial general intelligence (AGI) systems that can research AGI themselves, we have automated automation which will significantly accelerate AI's development. Dawes (2020) states that "This kind of asymptotic, accelerating acceleration means that even if the run-up to human-level AGI is quite long, the next major step to superintelligence[2] could be quite rapid". Furthermore, even if we are still very far away from developing AGI (which is highly probable), we should still consider the possibilities and risks. As Stephen Hawking once put it: "So, facing possible futures of incalculable benefits and risks, the experts are surely doing everything possible to ensure the best outcome, right? Wrong. If a superior alien civilization sent us a message saying, 'We'll arrive in a few decades,' would we just reply, 'OK, call us when you get here – we'll leave the lights on'? Probably not – but this is more or less what is happening with AI." (The Independent, 2014), which shows the need to consider the implications, benefits, and risks of AGI and superintelligence.

[1] `https://www.quanticdream.com/en/detroit-become-human`

[2] Superintelligence refers to AI that is far more intelligent than humans in every aspect, whereas AGI refers to intelligence on approximate human level.

Even if the probability of self-awareness in AI systems is low in the near future, it cannot be ruled out entirely. Therefore, this essay will focus on discussing whether, and under what conditions, we should grant (human) rights to these autonomous self-aware entities. Addressing these questions proactively may help avoid future ethical and societal tensions arising from uncertainty about their moral and legal status.

2.3 Affected Actors

It is clear that granting rights to self-aware AI systems has major consequences for these systems themselves. It is, however, difficult to determine *when* rights should be granted. Dawes (2020) argues that consciousness should entail rights, in line with how the Great Apes Project of 1994 granted apes basic human rights. New Zealand's Animal Welfare Amendment Act of 2015 shows similarities and made New Zealand the first country in the world with legislation recognizing animals as "'sentient' beings to whom we have duties" (Dawes, 2020).

The main difficulty here is defining consciousness. Some people argue that consciousness is a property of biological organisms, whereas others argue that we can fully compute consciousness, implying AI could do it too (Dawes, 2020). Furthermore, even if AI could possess consciousness, how do you determine if it has it, and is not merely simulating it? At the time of writing, this is still an open problem. If, however, AI turns out to have developed consciousness, and we dismiss it as merely a simulation of consciousness and continue to use them as tools, that could potentially result in slavery (Schneider, cited in Dawes (2020)).

Importantly, not only the AI systems are impacted. Granting AI systems human rights impacts humans as well, and has impact on society as a whole. As discussed in a blog post by Saravanan (2024), granting human rights to non-human entities could "alter human relationships, redefine societal roles, and challenge existing social structures". He argues that AGI could drive innovation, but could also lead to job displacement and economic inequalities, affecting working families. Whereas using AI as a tool still requires human operators, AI systems that have the same workers' rights as their human counterparts could compete with humans for jobs, causing friction.

While these questions may appear novel due to the involvement of AI, they closely resemble existing ethical debates surrounding moral status and rights, such as those concerning animals or other non-human entities. However, the scale, autonomy, and potential intelligence of AI systems introduce fundamentally new dimensions to these long-standing issues.

3 Solution

3.1 Ethics

The question of human rights for AI systems is inherently connected towards the question of the moral status of these systems. Based on this moral status, we can determine if something is morally wrong and why it would be wrong. An example is given by Sinnott-Armstrong and Conitzer (2021): it is morally wrong to destroy someone else's car, because the owner has the moral right to not have their property destroyed. You are wronged, not your car.

At the same time, it is morally wrong to burn your own pet alive, not because you are its owner, but because the pet itself has the moral right to not be caused harm.

The ethical question thus becomes: how to define moral status? Harman (2003) (cited by Sinnott-Armstrong and Conitzer (2021)) claims that moral status is binary. An entity simply has moral status or lacks moral status. Although she acknowledges the fact a human is more important than for example an anaconda, she related this to the fact that the human will be able to remember the pain for longer and will therefore suffer more than the snake.

This view, however, has some significant limitations, as discussed by Sinnott-Armstrong and Conitzer (2021), who propose a different philosophy. In their view, moral status is to be defined in terms of strength and breadth. Strength refers to how strong a certain moral aspect applies to a certain entity. For example, imagine a situation where you must kill either a human or a spider. It would be morally wrong to kill the human. This implies that the human's moral right not to be killed is stronger than the spider's. Breadth refers to how many moral rules, rights, reasons, and wrongs apply to a certain entity. A baby, for example has a right not to be killed or harmed, while it does not have the right of freedom. Adults do have this right of freedom, implying that they have a broader moral status.

Sinnott-Armstrong and Conitzer (2021) claim that it is not enough to state that an entity has moral status, but it is even more important to state why it does. This reason would form the basis for its moral status, right, and protection. To formulate this reason, one can make use of two principles as set out by Bostrom and Yudkowsky (2014):

> *Principle of Substrate Non-Discrimination:* If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.
>
> *Principle of Ontogeny Non-Discrimination*: If two beings have the same functionality and the same conscious experience, and differ only in how they came into existence, then they have the same moral status.

Although these principles can guide the formulation of the explanation for moral status, they do not yet explain what its fundament is. A commonly used basis is sentience, but this does not fully align with our case of AI robots, as they may be self-aware, but not able to experience feelings (Sinnott-Armstrong and Conitzer, 2021). A more appropriate approach is to ground moral status in the properties of the entity itself, and to determine the extent to which it applies based on those properties. In other words, using properties of the entity to determine how broad their moral status is (Sinnott-Armstrong and Conitzer, 2021). As an example, take Frank from Section 1. Let us assume that Frank cannot experience pain. He therefore would have no moral right not to be caused pain, whereas granting the right of freedom would be applicable.

Sinnott-Armstrong and Conitzer (2021) further argue that advanced AI systems could, in theory, develop properties such as intelligence, consciousness, free will, or moral understanding. While the existence of such properties remains speculative, their possibility reinforces the importance of grounding moral status in the properties of an entity, rather than in its biological origin.

3.2 Society

As mentioned in Section 2, not only AI systems are impacted by the question of moral rights for AI, but human stakeholders as well. When AI rights clash with human rights, this can create friction, as discussed by Saravanan (2024). As an example, take the scenario from Section 1. If Frank is allowed to leave at 5PM, and receives rights similar to his human colleagues, this would place him on equal footing. Although this may be considered more ethically consistent, it could also lead to increased competition in the labour market, as AI systems would no longer be treated as tools, but as agents with similar rights and positions as human workers. This friction is shown in the *Detroit: Become Human* game as well, which depicts a world with widespread unemployment and protests against Androids (see Figure 1).

Figure 1: Representation of societal resistance and unemployment due to Android integration in *Detroit: Become Human*.

It is therefore crucial to take the perspective of human stakeholders into account too, with the goal of decreasing this friction. One approach to addressing this tension is to apply the moral status framework introduced in Section 3.1. While we have already discussed how the breadth of moral status can be defined, we have yet to define how the depth of AI systems should be positioned. Sinnott-Armstrong and Conitzer (2021) discuss this in less detail compared to the breadth, but do provide us with some pointers. We already saw the example of killing a fly or killing a human, and stated that killing a human would be morally wrong. Thought experiments, adapting the original trolley problem, have started emerging and are focusing on whether you should put a human at risk to save a larger number of seemingly conscious AI systems (Dawes, 2020). This is a deeply moral question, which is impossible to answer until we can reliably determine consciousness.

However, such uncertainty does not imply that all entities should be treated as morally equivalent. In the absence of clear evidence that AI systems possess consciousness or the capacity for suffering, it is reasonable to assign them a more limited moral status in terms of strength compared to humans.

From a precautionary perspective, this suggests that while AI systems may warrant certain rights based on their properties, these rights should remain weaker than those of humans, ensuring that human interests are prioritized in cases of conflict.

A second, complementary approach to reduce friction and potential conflict between AI systems and humans is to incorporate Value Sensitive Design (VSD) (Friedman, 1996) into

the design of AI systems wherever possible. VSD aims to systematically integrate human values into the design process, which may increase the likelihood that AI systems remain aligned with these values. In the context of potentially advanced or self-aware AI, such alignment could help mitigate conflicts arising from differing interests between humans and AI systems.

3.3 Regulation

Current legislation already regulates the use of artificial intelligence. The European Union's AI Act (European Union, 2024), for example, adopts a risk-based approach, in which systems classified as high-risk are subject to stricter requirements. However, this regulatory framework remains fundamentally human-centric and does not account for the perspective of AI systems as potential moral subjects (Saravanan, 2024).

It is therefore important to develop the legislation that also considers the perspective of these self-aware AI systems. A possible starting point would be to translate the framework of Sinnott-Armstrong and Conitzer (2021) into legally binding rights, taking into account that the strength of AI's moral status should not exceed human's moral status, unless there is clear evidence that such systems possess consciousness.

To increase the general applicability of such frameworks, legislation could be extended beyond a strictly AI-specific focus. Instead, a more unified approach could be adopted that encompasses humans, animals, and AI systems, allowing them to be addressed within a single framework while still accounting for differences in moral status. Additionally, such an approach would enable legislators to proactively develop regulation, without being constrained to a potentially distant future scenario in which advanced AI systems emerge.

3.4 Normative reflection

In the previous sections, the ethical, societal, and legal perspectives on rights for AI systems have been examined. Although we are still far from developing AGI or self-aware systems, it is important to take these perspectives seriously. History has shown that regulation is often developed only after technologies have already had a significant real-world impact, as illustrated by recent developments such as the EU AI Act.

Developing a unified framework that encompasses humans, animals, and AI systems would allow such legislation to be debated and shaped proactively. This may help reduce future friction and conflict between humans and AI systems, while ensuring that legal frameworks are developed with careful consideration of all relevant perspectives, rather than under the pressure of reacting to unexpected technological developments.

4 Conclusion

In this essay, the question of whether AI systems can have rights has been examined from an ethical, societal, and legal perspective. A crucial issue is how moral status is defined, and which rights follow from it. Sinnott-Armstrong and Conitzer (2021) provide a useful

framework by grounding moral status in the properties of an entity, from which corresponding moral rights can be derived.

Current legislation, including the EU AI Act, does not yet account for these considerations and remains largely human-centric. The proposed approach of translating the framework of Sinnott-Armstrong and Conitzer (2021) into legislation, combined with the use of Value Sensitive Design where applicable, offers a promising direction for addressing these challenges. It remains essential to consider both human and AI perspectives when developing such legislation, while explicitly accounting for differences in the strength of moral status.

References

Nick Bostrom and Eliezer Yudkowsky. *The ethics of artificial intelligence*, page 316–334. Cambridge University Press, 2014.

James Dawes. Speculative human rights: Artificial intelligence and the future of the human. *Human Rights Quarterly*, 42(3):pp. 573–593, 2020. ISSN 02750392, 1085794X. URL https://www.jstor.org/stable/26947436.

European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj.

Batya Friedman. Value-sensitive design. *Interactions*, 3(6):16–23, 1996.

Elizabeth Harman. *Moral status*. PhD thesis, Massachusetts Institute of Technology, 2003.

Avinash Saravanan. Should general AI be given human rights?, 2024. URL https://medium.com/@asarav/should-general-ai-be-given-human-rights-dc4443da53c6. Accessed: 2026-04-14.

Walter Sinnott-Armstrong and Vincent Conitzer. How much moral status could artificial intelligence ever achieve? In *Rethinking Moral Status*. Oxford University Press, 08 2021.

The Independent. Stephen Hawking: Transcendence looks at the implications of artificial intelligence, but are we taking AI seriously enough?, 2014. URL https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html. Accessed: 2026-04-07.