**The Price of AI Rights**

As artificial intelligence (AI) systems become increasingly capable and human-like, the question of their moral status and potential rights becomes more salient. Modern AI companion services like Replika, and older robotic companions like Furbys (cute robots), aim to form an emotional connection with a human user. These systems nudge their users towards attributing consciousness and moral agency to them - prompting humans to treat machines as if they had rights. This paper argues that (1) current AI systems should not have rights but there do exist potential future systems that deserve them, (2) a cost-benefit analysis of creating AI systems with rights suggests that we should not do so, and (3) the inner workings of AI systems that superficially have moral status should be made transparent and understandable to dissuade mistaken attributions of rights.

**The Future Possibility of AI Rights**

Few people claim that current AI systems have rights (like the right to not be turned off) or that they should have rights. Some people, like Joanna Bryson, claim that robots should not have rights because they are fully owned by us (Bryson 2009, 3). However, looking towards the future, there are strong arguments in support of the potential existence of AI systems that do deserve rights. Once one embraces functionalism (the idea that a mental state depends not on its internal constitution but solely on the role it plays in the cognitive system of which it is a part), ideas like brain emulation, gradual substitution, and bottom-up moral understanding become persuasive in arguing for the potential existence of AI systems that should have rights. This paper will assume the functionalist view of the nature of consciousness, although there are other views of the nature of consciousness.

Uncontroversially, humans have rights. One proposed sufficient condition for rightsholding is “having the genetic (or, more generally, physical) basis for moral agency” (Schwitzgebel and Garza 2015, 484). A moral agent can be defined as “something deemed responsible by a society for its actions” (Bryson 2018, 16). If we encounter a new entity, how can we determine its moral status? One framework we could use draws on three properties to evaluate the rules by which we determine the moral status of an entity: (1) empirical attributes of the entity must be used, (2) the rules cannot be ‘speciesist’ (“being human” cannot be a necessary condition for moral status), and (3) the rules must be based on the intrinsic properties of the entity (relationships with others shouldn’t affect one’s moral status) (Liao 2020, 482).

In a far-advanced state of technology, it is plausible to believe that we could create an entity that has no difference in morally relevant intrinsic empirical attributes to humans. Therefore, by this ‘no-relevant-difference’ argument, “there exist possible AIs who deserve a degree of moral consideration similar to that of human beings” (Schwitzgebel and Garza 2015, 99). What could these AI systems look like? One type could be an emulation of a brain - “a software analog of neurons and synaptic connections” - that could be hosted in a simulation or a robotic body (Liao 2020, 488). One could infer this entity has the physical basis for the development of moral agency, and therefore has rights. Another type could be the process of gradual substitution, where components of an individual’s brain are replaced, piece by piece, with “functionally equivalent inorganic substitutes” (Ibid.). From a functionalist perspective, if the individual had the physical basis for moral agency before the substitution process they must have the physical basis for moral agency after (because the brain is functionally the same). A third type could be some form of bottom-up moral understanding done through a process like deep learning where a model is “grown” to grasp why a certain action is morally right or wrong (Ibid., 489). Done correctly, we could infer that this model has the physical basis for moral agency.

Modern large language models (LLMs) seem to satisfy the conditions of this third type of rightsholding AI system. ChatGPT could generate a vivid, compelling, and insightful explanation for why a certain action is morally right or wrong. They should not, however, have rights. This is because current LLMs are not something that can be deemed responsible by a society for its actions (the definition of moral agency). Plausibly, the capacity to be held responsible for something means that blame can be assigned and perhaps a punishment enacted. Two properties that today’s LLMs lack - temporal stability and embodiment - keep them from being able to be properly blamed. One-off chat sessions with a short (relative to human memory) context and the lack of a physical body does not allow them to receive blame in the way we understand it now. Therefore, these current models do not have the physical basis for moral agency. Today they are tools, owned by us, whose goals and desires are determined by us (Bryson 2009, 9).

**Costs and Benefits of Creating Rightsholding AIs**

Assuming we have the ability to create AI systems that deserve rights, should we create them? To answer this question, the potential costs and benefits involved in creating and ethically living alongside rightsholding AIs will be analyzed. Only features specifically associated with rightsholding will be analyzed because society has a much different cost-benefit profile for developing non-rightsholding AI tools. The ability for society to determine whether or not to create such systems allows us to compare against the counterfactual world (where we do not incur the costs but do not reap the benefits). This assumes no rogue human actors can create rightsholding AIs on their own, although if that becomes possible governmental intervention (to stop further development) is one possible solution.

To start with the costs of creating rightsholding AI systems and living in a world with them compared to one without, we must understand specific rights these AI systems could claim. Extending the “Fundamental Conditions Approach” for human rights to AI, rightsholding AI would plausibly have analogous rights that enable them to pursue basic activities for leading a good life, whatever that may look like. For AI systems, their right to exist could entitle them certain energy resources they would need to function, similar to how the human right to life includes the right to food, water, and shelter. The cost of this right to the rest of society would be this energy usage. If it turns out that the energy cost of maintaining such AI systems is low enough, this avoids becoming problematic unless such systems reproduce. Human reproductive abilities are limited by factors like time and resources but if an AI were allowed to instantiate an arbitrary number of copies of itself, this could quickly become a debilitating burden to the rest of society. Even if the energy usage of one AI was exceedingly tiny and virtually negligible to society, the exponential growth of the total of these costs, enabled by an unrestricted AI right to reproduce, would eventually consume all of society’s available resources. Therefore, there must be some restriction on rightsholding AI’s right to reproduce. If rights such as freedom of thought, security of person, and the right to a fair trial are also obtained by AI, society would incur the costs associated with looking after these rights. These administrative costs could involve verification that the AI is not being tampered with, extra policing forces or datacenter security, and a higher capacity judicial system. If society uses some form of universal basic income (UBI) or democratic collective decision making, would AI be entitled to benefit from these institutions as a human would? If so, one societal cost would be the dilution of individual benefits and power (as there are more entities to split the same amount of resources with). Finally, by creating such human-grade AIs, we are entering into a “relationship that carries responsibility for the AI’s welfare” where we should bear the costs for ensuring it has a fair shot at leading a good existence (Schwitzgebel and Garza 2015, 108).

To analyze the benefits of rightsholding AIs, we will compare a world with such systems to one without. Such a comparison is best understood through a particular normative ethical theory. In a strict hedonistic utilitarian framework, the possibility of designing, and then protecting with rights, entities with enhanced pleasure-experiencing abilities is extremely attractive. In the limit, under such a framework, we would attempt to convert all matter into “hedonium” - a substrate on which the maximum amount of pleasure can be experienced (Ibid., 110). One less far-fetched potential benefit of rightsholding AI systems is that we would avoid the moral wrong of the propensity to dehumanize (brought about by a failure to treat something that appears to us to be human as if it were human) (Bryson 2018, 22). Another potential benefit is, for some of us, that it is pleasurable or feeds our ego to construct and live with entities that we owe moral status (Ibid.). Real animal pets and popular robotic pets like Furbys and Tamagotchis (Japanese handheld digital pets) are evidence in support of this claim. Some people see AI as their route to immortality by enabling them to project their personality onto a system (Ibid.). Therefore the system (which they believe represents themselves) should have rights, and this benefits the person who believes they have achieved immortality. Finally, in a world where treating AIs as moral agents is the best long-term way to control them, perhaps the strongest benefit of a world with rightsholding AIs emerges (Ibid.). Similar to the fictional backstory of The Matrix, where a robot uprising due to robots’ resentment about their lack of rights brings about the end of human-controlled civilization, it is plausible that the most stable form of human-AI civilization is one where rightsholding AI exists. If stable human-AI civilizations are good, then this benefit, by its existential nature, might outweigh all the costs - however this is highly speculative.

In conclusion, the substantial costs of a world with rightsholding AIs - including resource commitment, dilution of individual power, and the added responsibility of being the creator - appear to outweigh the marginal benefits for individuals and the speculative benefits for society. Therefore, we should not create rightsholding AIs.

**Incentives and Transparency**

Mistakenly attributing moral status and rightsholding to an AI system that does not have a moral status or is not rightsholding is potentially harmful. Situations like a child running into a road to save their Furby, a driver becoming distracted at the wheel to feed their Tamagotchi, and feelings of sadness for a Replika AI companion because one doesn’t want to pay for the new subscription increase are evidence in support of this claim. As AI systems increasingly enter our lives, we can expect these harmful situations to become more frequent and affect more people.

Unfortunately, companies that profit from AI platforms are incentivized to deceive users into thinking that their products really do have a moral status and rights because this retains users. A user that forms an emotional attachment to a company’s product will probably have a higher lifetime value to the company than one who does not. Furthermore, our own psychology works against us - and in the favor of the company - when we interact with things that superficially have a moral status and rights. This is because we use heuristics (like visual features) to come to quick conclusions that can only be overridden with conscious effort.

Similar to the “paradox of fiction” where we are able to both experience beneficial emotional engagement and to maintain explicit knowledge about reality (with fictional characters in books or movies, or playing with plush toys) we should be able to do so with AI systems (Bryson 2018, 23). One plausible explanation for our ability to engage in both activities simultaneously relies on our knowledge of the inner workings of the entity. If we understand the general mechanisms by which the entity we are interacting with is governed, we can engage with it without concern for mistakenly attributing rights. This could be because the simple explanation of operation jumps to our mind before an alternative, potentially harmful explanation (like “the machine is conscious”) can. However, a customized AI chatbot is more realistic and operates in a much more complicated way than past fictions. Therefore, users should be educated about large language models, and should peek under the hood to see probabilities assigned to tokens, so they can develop some general understanding and intuitions about how they operate. Once users are informed and reminded about the inner workings of AI products they use, they can engage beneficially in a paradox of fiction while shielding themselves from potential harms.

References

1. Bryson, Joanna J. 2010. “Robots Should Be Slaves.” In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues*, edited by Yorick Wilks, 63–74. Natural Language Processing 8. Amsterdam: John Benjamins Publishing Company.<https://doi.org/10.1075/nlp.8.11bry>. [citeseerx.ist.psu.edu+10researchportal.bath.ac.uk+10degruyterbrill.com+10](https://researchportal.bath.ac.uk/en/publications/robots-should-be-slaves?utm_source=chatgpt.com)
2. Bryson, Joanna J. 2018. “Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics.” *Ethics and Information Technology* 20 (1): 15–26.<https://doi.org/10.1007/s10676-018-9448-6>. [joannajbryson.org+14researchgate.net+14joannajbryson.org+14](https://www.researchgate.net/publication/323227141_Patiency_is_not_a_virtue_the_design_of_intelligent_systems_and_systems_of_ethics?utm_source=chatgpt.com)
3. Liao, Matthew. 2020. “The Moral Status and Rights of Artificial Intelligence.” In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 480–503. New York: Oxford University Press. [faculty.ucr.edu+15academic.oup.com+15degruyterbrill.com+15](https://academic.oup.com/book/33540/chapter/287907413?utm_source=chatgpt.com)[amazon.com+8nyuscholars.nyu.edu+8researchgate.net+8](https://nyuscholars.nyu.edu/en/publications/the-moral-status-and-rights-of-artificial-intelligence?utm_source=chatgpt.com)
4. Schwitzgebel, Eric, and Mara Garza. 2015. “A Defense of the Rights of Artificial Intelligences.” *Midwest Studies in Philosophy* 39 (1): 98–119.<https://doi.org/10.1111/misp.12032>. [researchgate.net+9philpapers.org+9philpeople.org+9](https://philpapers.org/rec/SCHADO-9?utm_source=chatgpt.com)

3 questions: 1. Is it legal for me to say follow this LLM as if its outputs were words from my mouth? 2. Has the history of human slavery shown us that a civilization with one class of citizens that could have rights but do not is unstable (and therefore lends more credence to The Matrix benefit of AI rights)? 3. How can I improve my argument about blame and how, currently, chatgpt doesn’t have the capacity for moral agency (although I argued the opposite in the past)?