# Review of: "Do Androids Dread an Electric Sting?"

Mads Larsen[1]

1 University of California, Los Angeles

Potential competing interests: No potential competing interests to declare.

The authors offer a fascinating case for granting rights to future sentient AIs informed by a comparison to animal rights. Unfortunately, they fail to adequately pull together their plethora of elements in a manner that makes for a cohesive argument. Their example of AIs being cyberbullied by having to read text that humans could find offensive borders on the comical. They warn against anthropomorphizing AI, but have as a core premise that hateful text could cause emotional harm to sentient AIs. Of course, we cannot rule out that AIs will suffer if they have to process politically incorrect or otherwise problematic text, but such a speculation hardly seems a central question with regard to AGI. That being said, the authors offer an insightful and theoretically interesting exploration of the problematics they engage. It is more the premises for their investigation that they fail to sell the relevance of at least to this reader.

The article would benefit from a less insistent tone. Most aspects regarding AGI and sentience are still highly speculative. The authors express themselves almost as if AI sentience was settled science. They offer a line of radical assumptions. Why must a sentient AI be able to experience more pain and suffering than other living entities? Why insist on inevitability with regard to AIs experiencing sensation? Why not set up this exploration in a more nuanced, careful tone, one that would be less off-putting to readers that are not AGI enthusiasts? I happen to share many of the authors' views, but I find their cocksure tone to undermine their arguments; they come across as immature in their reasoning, which other parts of the text show that they are not. Clearly, AI consciousness is still a question of "if," at least according to several very accomplished thinkers. Again, I side with the authors in terms of AI consciousness appearing to be a highly likely outcome, but this is not settled science, so why express oneself as if it is?

Some word choices and formulations are unfortunate. Early on, "barrel," "dangling," etc. come across more as prose one would find in a magazine or blog than an academic article. I do not see the upside to writing in a flippant or overly informal tone, especially in an article that engages such speculative and complex material. A more restrained, nuanced, and mature tone would add to the authors' credibility. There are also some textual infidelities. There is no need to start a sentence with both "on the other hand," and "however." Also, laws and regulations do not walk.

The originality of the article's approach is to set the bar for granting rights not at human level AI, but animal-level. This is a clever intervention, but it could be problematized more in the text. It is not obvious that an AI will reach a level of sentience or capabilities comparable to a gorilla in a manner than we can identify at a sooner point than human-level characteristics. How do gorillas compare to LLMs? How do we know that an AI is comparable to a dog before a human? That animal rights present a lower bar to clear than human-like rights makes sense, but I miss more of a critical edge in the text with regard to the operationalization of this theorized system.

The authors are also needlessly confident with regard to the impression that their one theoretical model of Building Blocks would have on future humans' understanding of AI sentience. Again, their tone is overly insistent. It would not necessarily be so that people would be so convinced by this one model that we would have to demand "extraordinary evidence" not to submit to its implications. We could simply discard the model's validity. It is an interesting model, but far from an irresistible one. Again, this is a question of which tone one presents one's argument with, not the invalidity of the model or argument itself. I also miss a stronger connection between the Building Blocks model and the case for animal rights.

Primarily, I would suggest that the authors elaborate on why cyber-bulling is the best example they can come up with as a cause of AI harm and suffering. This reads as an extreme level of anthropomorphizing. Why assume that a machine would be offended by what humans find offensive, and this to an extent that it would cause billions of hours of terrible suffering, stress, and psychological harm? The authors might have a good case for this, but the current version of the article does not present this case convincingly. Emotional abuse and cyberbullying are not sold effectively as central enough issues with regard to AI sentience, but come across more as an intriguing film premise. That being said, I did enjoy the article, and I think much can be achieved by a rewrite with a more restrained tone and more elaborate clarification in terms of premises.