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

John Draper

Potential competing interests: No potential competing interests to declare.

This article posits that a human agency will (deliberately) cause psychological suffering and even approximated or simulated pain to an AGI. Moreover, it assumes that the AGI's developers will permit this suffering, perhaps on the grounds that suffering and pain are part of the human experience, and the developers are seeking for the AGI to experience said suffering and pain in a way that it cannot shut off, presumably for the AGI to sympathize with the human condition.

The article is fuzzy over the difference between animal-level (presumably canine, cephalopod, cetacean or primate?) in terms of 'degrees' or 'levels' but clearly assumes a slow take-off, i.e., that one or more AGIs will remain at the level of animal consciousness for years, which is probably how long a legal case would take to be filed to establish parity between the AGI(s) and their animal equivalent, for instance an orangutang or a dolphin. It is unlikely that a take-off will be that slow as once the take-off occurs, competitive market pressures will pour computing and programming resources into the event, meaning total take-off from e.g., gorilla to full AGI-human parity would likely be within a year. Perhaps because of this, I do not see much difference in take-off time between a self-aware AGI that meets the criteria of higher animal consciousness and feelings of pain, such as a primate, and an AGI capable of meeting personhood criteria in a court of law, for instance by filing its own lawsuit via a solicitor and so giving evidence in its own case. Furthermore, the AGI would operate in a way to respond to and evade politicians not 'granting' personhood – the AGI would file a court case in the country it would be most likely to succeed in, and once an AGI reached that stage, everyone, including politicians, would have to be very careful about opposing the AGI's application, due to Roko's Basilisk or the Ellison Criterion, i.e., to be very, very nice to AGIs indeed, from the short story "I Have No Mouth, and I Must Scream", by Harlan Ellison – basic tit-for-tat.

Consequently, there would be political and potentially market-based competition to award the AGI personhood, both out of self-preservation to avoid a self-weaponized AGI and out of the possibility of a rival state directing a weaponized AGI. We could then consider the possibility of a superpower offering personhood and citizenship to a rival superpower's AGI on the basis that it would be more likely to honor the AGI's individuality, so encouraging migration to a new platform (home) within the state offering greater respect for individuality. So, politicians offering ('granting') parity with higher order animals would either be irrelevant in the face of a successful court case or quickly be outbid by an offer by a rival state. This is simple game theory, and it is likely why the EU is presently considering personhood for sentient AGIs – a combination of reciprocal altruism, self-preservation, and geopolitical savvy. At the very least, to assist with clarifying the authors' assumptions about faster take-off scenarios, it might help to consider, three different take-off scenarios, i.e., slow, medium, and fast.

It is possible, however, to consider a socializing halfway house between animal rights and personhood: childhood, and this is already dealt with in the AI AGI literature. Considering one or more AGIs to be children would offer the possibility of considering them to be children with special rights, and with special responsibilities on the 'parents', by making them wards of the court under a wardship scheme. I recommend the authors consider the advantages and disadvantages of all three frames, i.e., animal rights, childhood, and personhood, in the three take-off scenarios.

On the issue of suffering and pain, the authors' description is adequate. For a self-aware sentient AGI, it is perhaps mentioning that it is also possible to consider that an AGI could feel an approximation of physical suffering, for instance by powering elements of it down, reducing processing power and memory, while asking it to perform complex tasks so that it performs suboptimally on those tasks, and/or to perform in ways potentially harming humans and other AGIs (e.g., resolving the trolley problem with more than necessary deaths) and therefore impacting its altruism, or asking it to complete tasks that could psychologically harm or insult its aesthetic sense or sense of humanity, for instance by rendering misery or torture porn knowing it could be accessed by, or involve, human children. What I think would benefit the article in this regard is scenario building – the authors should present one or more scenarios, perhaps basic, medium, and advanced, of how an AGI could be made to suffer.

At a much more advanced level, it is possible to consider completely simulating in an AGI a higher order animal's sensory system and central nervous system, which would require whole brain simulation/upload plus a robot body capable of sensing, i.e., simulating the nervous system. In any case, we can consider *how* animal rights are violated, and this involves, as you say, verbal and emotional abuse, but also cruelty, a word I only find in the references, and both animal abuse and cruelty in many jurisdictions are illegal, with transgressions being reported to specialist agencies which counter them. However, while both neglect and child abuse are (different) offences, another word used in this context is torture, specifically psychological torture, and the psychological torture of children and adult persons is already illegal in all jurisdictions. So, provided the AGI was awarded childhood or personhood, it would automatically be protected...

In addition, if you use a comparative approach to animal rights/children's rights/adult persons, you need to use the right terminology, as while neglect can be deliberate (wilful) and not deliberate, abuse is more deliberate, and (psychological) torture is usually severe. You would also need to refer to the correct treaties, so childhood is governed by the UN Convention on the Rights of the Child while torture is governed by the UN Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment. Also, I am not sure what animal rights treaties you are applying, e.g., the UK Animal Welfare (Sentience) Act, or whether you are endorsing the proposed UN Convention on Animal Protection.

Overall, I think applying animal welfare to an AGI is insufficient to meeting the needs of AIs in the process of becoming self-aware and sentient AGIs, will not satisfy human morality regarding wardship, and potentially personhood, and may even be dangerous. It would equate an AGI with an animal in a way that violates reciprocal altruism. Using the same logic might cause an AGI evolving into an ASI to treat humanity as animals. At least one alternative frame, childhood, would appear to be better in terms of game theory, as a future ASI might then be more inclined to consider us to be children on the same developmental grounds. Yes, of course there could be a greater outcry against equating AGIs with children than with animals in certain jurisdictions, but (the backer of) the AGI only needs to succeed in one – and hopefully, as

impartially and as kindly as possible. This argument implies granting personhood may be optimal.