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

Louis Irwin[1]

1 University of Texas at El Paso

Potential competing interests: No potential competing interests to declare.

ALTERNATIVE SCENARIO FOR EMERGENCE OF SENTIENT ARTIFICIAL INTELLIGENCE

Overview

As the extent of artificial intelligence (AI) rapidly grows, spurred by commercial competition, the need to anticipate the possibility of its negative as well as positive consequences increases. One such potential result is the possibility that AI entities could become subjectively aware, or conscious. This article poses such a possibility and suggests how it should be viewed and dealt with by humans. The topic is important and the article provides a thoughtful starting point for a discussion that society needs to undertake.

A summary of the authors' position, as I understand it, is as follows:

AI systems are approaching the capability of enabling consciousness. When AI systems do become conscious, they will then be capable of pain and suffering. Humans have an ethical obligation to prevent this. The way to do so is to adopt the same legal standards and laws that are applied to animals deemed worthy of welfare consideration, which are those that can experience negative affect (with emphasis on pain and suffering). Welfare consideration means affording them protection from harmful experiences and the freedom to express 'natural' behavior.

The authors' argument is well presented and logically coherent. However, I believe there are weaknesses in the argument at several points that lead me to some different assumptions and conclusions. I will elaborate on those points of disagreement, then suggest an alternative scenario for dealing with the prospect of sentience in artificially intelligent entities.

Points of Disagreement

AI systems are approaching the capability of enabling consciousness.

There is no question that AI systems are rapidly gaining in complexity and sophistication, to the point that their linguistic output and manual proficiency (of machines and robots) display intelligent-like capabilities suggestive of conscious control. But since the mechanisms that enable consciousness are unknown in humans (the only species of which we have direct knowledge of consciousness), much less in other animals, it is premature to assume that any AI entity, regardless of how sophisticated, can engender phenomenological experience. While I agree that animal consciousness may have arisen by different trajectories and can be instantiated by different neural architectures [1-3], absent knowledge of what

any of those trajectories are or have been makes any prediction about the likelihood that an AI system could become conscious as mere speculation.

In fact, three features of what we now believe about consciousness seem incompatible with its likely manifestation in mere algorithms or machines. First, neurobiologists increasingly are adopting the view that phenomenological experience is *embodied*, or resident in and related to the whole body of an organism [4-6]. This means, among other things, that subjective awareness is experienced in different parts of its physical body, with reference to the function of its different body parts. Secondly, consciousness is inevitably *emplaced,* or situated in and related to the spatial context within which it operates [7, 8]. Thirdly, compelling arguments have been made that movement and the need to monitor and coordinate the action of a body in motion was the driving force and is the continuing utility of consciousness [8, 9]. Neither motionless computers nor algorithms or virtual entities like chatbots have these three features. Mobile androids could be engineered to have them to some degree but that would require an intentional act by their designers and fabricators.

Thus, it is certainly conceivable that some AI systems could become conscious if deliberately engineered to be so, but it is by no means inevitable.

When AI systems do become conscious, they will then be capable of pain and suffering.

Throughout the article, the authors conflate consciousness with pain and suffering. There are at least three reasons why this is unjustified. First, human consciousness is ordinarily present without pain and suffering, so while consciousness may be necessary for aversive feelings, it is certainly not sufficient. Secondly, pain has evolved in animals to promote the avoidance of tissue damage [10] — a decidedly biological problem. No analogous threat is obvious in AI algorithms or machines, so the emergence of pain as an accompaniment of consciousness carries no adaptive advantage in them. Of course, AI entities could be constructed to perceive and resist threats of various sorts, but this would require deliberate engineering. Thirdly, as the authors commendably point out, pain as we know it is an explicit sensory perception mediated by a specific neurological substrate, involving unique nociceptors and dedicated neural pathways that project to precise brain regions. The authors assume that analogous systems could be engineered in AI systems and programmed to generate negative affective feelings in systems capable of phenomenological experience. But this would require deliberate engineering for the purpose of inflicting duress on the sentient AI entity. Absent the deliberate introduction of negative affect by design, there is no reason to assume that it would emerge unplanned just because an entity became artificially intelligent, even to a high degree.

Humans have an ethical obligation to prevent this.

I personally agree, but would point out that philosophers and ethicists do not view ethical obligations as based on any one particular set of values: in the case at hand, the prevention of pain and suffering. Indeed, humans tolerate the infliction of pain and distress on animals for a number of reasons, ranging from raising them for food under stressful conditions, confining them as pets, experimenting on them for human benefit, destroying them for agricultural and ecological reasons, and recreational uses. Do we have any reason to suppose that public attitudes toward duress in AI entities, if shown to exist, would be faced with any fewer conflicting attitudes?

The way to prevent abuse of conscious AI entities is to adopt the same legal standards and laws that are applied to animals deemed worthy of welfare consideration, which are those that can experience negative affect (with emphasis on pain and suffering).

This is not a good idea. Pain perception and negative affect have ancient evolutionary roots and are embedded in the neural capabilities of many if not most animals [1, 11], so animal welfare laws amount to prohibitions against activities or circumstances that animals innately perceive as painful or stressful. But algorithms and AI machines have no innate sense of duress, unless that type of experience has been enabled in their creation. A better, simpler way to prevent abuse of an AI entity is to prohibit the construction and programing of negative affect into the system in the first place.

By analogy with laws pertaining to the welfare of animals, welfare consideration for AI entities would have to mean affording them protection from harmful experiences and the freedom to express 'natural' behavior.

This would be at best problematic. The AI entity would have to be able to signal that it is being harmed or potentially could be. Presumably, the detection of present or potential harm could be engineered into the design or program of the AI entity, but why would it be if the capacity for harmful experience were not engineered into the entity's program in the first place? And what would constitute the 'natural' behavior of an algorithm, robot, or machine?

An Alternative Scenario

Based on my assumptions, I think the following scenario for the emergence and control of sentient AI entities is more likely:

AI systems are approaching human-like information processing capabilities. It is possible but not inevitable that some AI systems will be engineered to experience some level of self-awareness. When that happens, the possibility must be considered that AI systems could further be engineered to have adverse feelings about physical input through specialized sensors (analogous to animal pain) or semantic input into their deep language systems that provokes negative affect (analogous to duress in animals). Humans have an ethical obligation to prevent this. The way to do so is by enacting laws that make it a crime to engineer those capabilities into any form of artificial intelligence.

Conclusion

Consciousness arose in animals to serve biological imperatives: environmental awareness and discrimination, distinguishing between the organism's internal and external environment, ability to move about in pursuit of prey or avoidance of predators; orientation of the organism's place in the environment and dynamic changes in that environment in real time. These necessities for survival evolved under the pressure of natural selection to become an inherent aspect of the organism's behavior. No such pressures affected the creation of computer algorithms, static computers, or robots, so there is no reason to assume that they would emerge as unintended consequences of artificially created intelligence.

The capacity for sentience in principle could be programmed and engineered into AI entities, but if a desirable goal of society (with which I concur) is to prevent the advent of a capacity for negative feelings — certainly including pain and

suffering — in such entities, the best way to prevent it is to prohibit its creation in the first place. Proactive laws for doing that, rather than retroactive laws that mimic animal welfare considerations will be much more likely to deter the emergence of machines capable of anything analogous to animal pain and suffering.

A Final Note

Both the article by Tait and Tan and the alternative scenario I have envisaged in this critique make assumptions about the future that have not yet become a reality. While I have given what I believe are sound biological reasons to support my scenario over theirs, we are both speculating about a future that is not yet upon us, so I cannot assert with confidence that my scenario is more likely than theirs. Tait and Tan deserve commendation for tackling a difficult subject in a thoughtful way, and instigating what I hope will be a fruitful ongoing discussion.

[1] Ginsburg S, Jablonka E. 2019. The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness Cambridge, MA: MIT Press.

[2] Irwin LN, Chittka L, Jablonka E, Mallatt J. 2022. Comparative animal consciousness. Front Syst Neurosci. 16:998421. doi: 10.3389/fnsys.2022.

[3] Mallatt J, Feinberg TE. 2021. Multiple routes to animal consciousness: constrained multiple realizability rather than modest identity theory. Front Psychol. 12:732336.

[4] Dreyfus HL. 1992. What Computers Still Can't Do: A Critique of Artificial Reason. Cambridge, MA: MIT Press.

[5] Adams F. 2010. Embodied cognition. Phenomenol Cognit Sci. 9:619-28.

[6] Merleau-Ponty M. 1945. Phénoménologie de la perception (Phenomenology of Perception). London: Routledge & Kegan Paul, 1965.

[7] Clark A. 1997. Being There: Putting Brain, Body, and World Together Again. Cambridge, MA: MIT Press;.

[8] Irwin LN, Irwin BA. 2020. Place and environment in the ongoing evolution of cognitive neuroscience J Cogn Neurosci. 32:1837-50.

[9] Sheets-Johnstone M. 1999. The Primacy of Movement. Amsterdam: John Benjamins Publishing.

[10] Melzack M, Fuchs PN. 1999. Pain, general. In: Adelman G, Smith BH, editors. Encyclopedia of Neuroscience. 2nd ed. Amsterdam: Elsevier; p. 1547-50.

[11] Feinberg TE, Mallatt JM. 2016. The Ancient Origins of Consciousness: How the Brain Created Experience. Cambridge, MA: MIT Press.