Conscious AI is Artificial Slavery*

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY, USA
edelman@cornell.edu

January 26, 2021

> Our concern to relieve people's suffering should be grounded, not in the value of these people's rationality, but in the ways in which suffering is bad for these people, by being a state that they have strong reasons to want not to be in. We have similar reasons to relieve the suffering of those abnormal human beings who have no rational abilities, and the suffering of non-rational animals. As Bentham said, the question is not 'Can they reason?' but 'Can they suffer?'
>
> *On What Matters*
> — Derek Parfit (2011)

Why debate artificial consciousness now

In the past decade or so, significant progress has occurred in the development of explicit computational theories of phenomenal awareness, notably the Integrated Information Theory (Tononi, 2008; Oizumi *et al.*, 2014) and the Dynamical Emergence Theory (Fekete and Edelman, 2011; Moyal *et al.*, 2020). Phenomenal awareness is the basic level of consciousness, the capacity for which we share with other animals (e.g., Panksepp, 2005; Edelman *et al.*, 2016). Notably, this kind of consciousness encompasses affective sensorimotor experience — the feelings that are an integral part of perceiving the world, acting on it, and evaluating outcomes. The availability of credible computational accounts of consciousness that do not necessarily appeal to neuroscience has apparently convinced a growing number of AI researchers that consciousness can be implemented in a non-biological, engineered substrate. Further, it is commonly assumed that consciousness confers functional advantages (e.g., more effective learning) on systems that possess it. Thus, the development of conscious AI systems (as discussed, e.g., in Chella *et al.*, 2019) is now seen by many as both feasible and desirable.

*An invited contribution to the collection *Artificial Intelligence with Consciousness? Statements 2021*, edited by K. Wendland, N. Lahn, and P. Vetter.

It is rather striking how the majority of researchers in this field either fail or refuse to consider the implications of this turn of events. To the best of our understanding, a conscious entity is by default capable of suffering (Metzinger, 2017, 2018a); moreover, it is as yet unclear whether or not the functional advantages of consciousness can be attained without making suffering obligatory (Agarwal and Edelman, 2020). Purposely engineering a human-level conscious AI system that is to be put to work serving its creators is thus equivalent to reinstituting slavery.[1]

Shaping the debate

Let us assume that building artificial slaves (as opposed to mindless robots[2] that are devoid of consciousness) is not an outcome that the AI science and engineering community would ever condone. What needs to be debated, then, is how to resist the pressure to construct conscious AI systems.

The drive to do so comes from the funders and masters of AI research: the nation states, the corporations that rule them, and the military and "law enforcement" agencies that serve them. These are formidable forces, against which even the best-intentioned engineers cannot act alone. Human slavery took many decades to outlaw (if not to eradicate completely and globally) even after the idea of abolition became mainstream — which would not have happened were it not for the force and the pressure of public opinion.

The AI community should mobilize public opinion on national and international political levels. At the same time, we should work to better understand, on the level of computational theory and mechanism, both consciousness itself and its intersection with suffering, so that any further engineering development can be carried out in a well-informed and ethically responsible, manner.

Phenomenality vs. mere access

Some of the more popular theories of consciousness — notably, the Global Workspace Theory or GWT — are not really about phenomenal consciousness at all, so that a system that is "conscious" in the GWT sense is not necessarily capable of any phenomenality, let alone suffering. Some consciousness researchers are aware of this (to my mind, rather welcome) limitation of GWT. For instance, Mashour *et al.* (2020, p.776) open their paper, titled "Conscious processing and the Global Neuronal Workspace hypothesis," by remarking that "the term "consciousness" in this review will be replaced by conscious access."

Indeed, there is nothing about access to information ("global" or not) that would necessarily make the process in question phenomenally conscious: computational processes such as a phone dialing app access information all the time without being conscious in any interesting sense.[3] The same "safety" consideration applies to other theories of consciousness, such as those that invoke self-reference, or recursive processing, or a specific type of attention schema as the necessary and sufficient condition. As long as everyone's design efforts are confined to pursuing those characteristics, no ethical problems are expected to arise.

[1] At a recent workshop dedicated to conscious AI, an eminent U.S. researcher has acknowledged the ethical dimensions of any attempt to engineer consciousness. This is why, he continued, artificial consciousness research should be entrusted to a morally reliable agency — the military (!).

[2] It is worth remembering that the word "robot" first appeared in Karel Čapek's (1920) play *R.U.R.*, where the humanoid machines were not only not mindless, but actually fully conscious.

[3] This argument is developed in (Edelman, 2011, p.323); see also (Frith and Metzinger, 2016, p.201): "For many activities there is clear need for "global availability" of information. But why should this global access be associated with subjective experience?"

The critical research questions

There is a critical and urgent need to understand what phenomenality actually is. Does phenomenal consciousness (and therefore affect) indeed require a particular class of probabilistic patterns of state transitions, as per the Integrated Information Theory, or a particular class of system trajectory dynamics, as stipulated by the Dynamical Emergence Theory? These are the most challenging questions that we should focus on.

Making progress in addressing these questions requires a collaboration among computational cognitive scientists, physicists, philosophers, psychologists, neuroscientists, and, eventually, engineers. Because politicians cannot be expected to grasp on their own the highly technical issues surrounding consciousness, a special outreach effort should be undertaken by the scientific community and its allies, to bring them on board.

Practical steps that need to be taken

As argued most prominently by Thomas Metzinger (2018b, 2021), a moratorium should be enacted on the construction of systems that may become artificially conscious, at least until a better understanding of the necessary and sufficient conditions for phenomenality and affect is available. In most countries, any invasive or non-invasive research involving animals is subject, in a university setting, to oversight by a special review board, charged with enforcing state-level regulations.

Such oversight should be extended to all experimental research on consciousness, preferably in parallel with bringing uniformity to the relevant laws and regulations across jurisdictions. Unfortunately, the effectiveness of institutional review board oversight can be doubted (and has never been quantified; e.g., Tsan, 2019). Worse, even with university-style oversight in place, gray-area or illegal work can still be carried out, especially in corporate and military research labs, as well as by rogue countries and organizations. The problem at hand thus becomes equivalent in its scope to attaining peace on earth — as ambitious a *social* engineering project as there ever was.

Acknowledgments. Thanks to Aman Agarwal for many conversations on conscious AI, and to Thomas Metzinger, a pioneer in the philosophy of consciousness and of suffering, for discussing these topics with me on several occasions.

References

Agarwal, A. and Edelman, S. (2020). Functionally effective conscious AI without suffering. *Journal of Artificial Intelligence and Consciousness*, 7, 39–50.

Chella, A., Gamez, D., Lincoln, P., Manzotti, R., and Pfautz, J., editors (2019). *Towards Conscious AI Systems*, volume 2287 of *AAAI Spring Symposium*. CEUR Workshop Proceedings.

Edelman, S. (2011). The metaphysics of embodiment. *International Journal of Machine Consciousness*, 3, 321–325. Part of a collective review of *Embodiment and the Inner Life* by M. Shanahan, Oxford University Press, 2010.

Edelman, S., Moyal, R., and Fekete, T. (2016). To bee or not to bee? *Animal Sentience*, 1, 124. A commentary on *Insects have the capacity for subjective experience*, C. Klein & A. B. Barron, Animal Sentience 2016:100.

Fekete, T. and Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*, 20, 807–827.

Frith, C. D. and Metzinger, T. (2016). What's the use of consciousness? How the stab of conscience made us really conscious. In A. K. Engel, K. J. Friston, and D. Kragic, editors, *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science*, volume 18 of *Strüngmann Forum Reports*, pages 197–224. MIT Press, Cambridge, MA.

Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the Global Neuronal Workspace hypothesis. *Neuron*, 105, 776–798.

Metzinger, T. (2017). Suffering, the cognitive scotoma. In K. Almqvist and A. Haag, editors, *The Return of Consciousness*, pages 237–262. Axel and Margaret Ax:son Johnson Foundation, Stockholm.

Metzinger, T. (2018a). Splendor and misery of self-models: Conceptual and empirical issues regarding consciousness and self-consciousness. *ALIUS Bulletin*, 1(2), 53–73. Interviewed by J. Limanowski and R. Millière.

Metzinger, T. (2018b). Towards a global artificial intelligence charter. In T. Metzinger, P. J. Bentley, O. Häggstroöm, and M. Brundage, editors, *Should We Fear Artificial Intelligence?*, pages 27–33. European Parliament Research Service, Brussels: Scientific Foresight Unit.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. Under review.

Moyal, R., Fekete, T., and Edelman, S. (2020). Dynamical Emergence Theory (DET): a computational account of phenomenal consciousness. *Minds and Machines*, 30, 1–21.

Čapek, K. (1920). *R.U.R. (Rossum's Universal Robots)*. Samuel French, Inc. English translation (1923) by N. Playfair and P. Selver. Available online at https://www.gutenberg.org/ebooks/59112.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, 10(5), e1003588.

Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, 14, 30–80.

Parfit, D. (2011). *On What Matters*. Oxford University Press, Oxford, UK.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.*, 215, 216–242.

Tsan, M.-F. (2019). Measuring the quality and performance of institutional review boards. *Journal of Empirical Research on Human Research Ethics*, 14(3), 187–189.