CEUR-WS.org/Vol-2287/paper1.pdf

Evolution of Conscious AI in the Hive: Outline of a Rationale and Framework for Study

David Sahner M.D.

Chief Scientific Officer, EigenMed

Abstract. The paper proposes a framework for investigating a potentially conscious AI system by considering an autonomous, multitask-capable, powerful, highly intelligent and adaptive system (AMPHIA) as one more likely to acquire consciousness and ethical principles through cooperative behavior, evolution, experience, training, and pedagogy.

1 Introduction

As artificial intelligence capabilities increase and machines start to exhibit characteristics that we previously thought to reside within the unique domain of humanity, the question of machine consciousness becomes pressing. For example, as machines become more intelligent, will they naturally become sentient? If so, what might be the implications for the safety and welfare of the inhabitants of future societies? How would consciousness affect their ability or willingness to discharge tasks? And what obligations might we owe to such creatures?

To begin to answer such questions in a secure and protected setting, prior to the unintended onset of consciousness in highly intelligent, powerful and autonomous systems, it is reasonable to consider efforts to facilitate evolution of machine consciousness under controlled and contained conditions, and to carefully study the effect it may have on both functional competencies and moral behavior. Such an approach also enables interventions at an early juncture in the emergence of consciousness that predispose to artificial phronesis [10]; that is, the quick and reliable instantiation of learned ethical codes in a social context under personal tutelage. In this way, two goals are achieved: (1) a proactive understanding of the potential effects of machine consciousness on the benefit-risk profile of highly intelligent autonomous systems, and (2) a greater likelihood that a sound moral compass will guide emergent behavior in such systems.

Although the ability of machines to "wake up" remains speculative, it can be hypothesized that lessons from the human evolution of consciousness can be leveraged to instill consciousness in machines. Although it is not possible to definitively cite an exact watershed moment marking the onset of modern human consciousness among *Homo sapiens* or its ancestors, it is likely that human consciousness evolved over time in an embodied, and cultural context, in which individual agents equipped with the necessary “wetware” needed to both compete and cooperate in exceedingly complex societies. Such an assumption is consistent with extensive observations made by Julian Jaynes in his work, *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, even if the neuroscientific basis for consciousness adduced in his theory lacks current support. Mimicking such an evolutionary trajectory within a condensed time frame in silico prior to the construction of physical robots that embody features of consciousness may be expedient. Apart from evolution, either virtual or physical, it may also be useful to initially leverage neuromorphic architecture, multimodal sensing, specific types of reinforcement learning, and current theories of human consciousness in our efforts to generate and study emergent phenomena in machines.

2 Autonomous, multitask-capable, powerful, highly intelligent, and adaptive system (AMPHIA)

The endowing of an autonomous, multitask-capable, powerful, highly (or super-) intelligent, and adaptive system (hereafter referred to as AMPHIA) with phenomenal consciousness may temper the risks to inhabitants of future societies. Several publications have made the point (see, e.g., [11; 12]) that consciousness may be required for empathy and moral-decision making.

As an alternative to fostering acquisition of cooperative behavior and ethical principles in a social framework through evolution, experience, training, and pedagogy, moral codes can be built into AI “from the ground up” at the operating system level. Formalized logic enabling automation of the doctrine of double-effect (i.e., allowing for the necessity of producing unavoidable harm for the greater good) has been pioneered by Bringsjord [4] but has yet to be advanced to a level that enables the handling of a multitude of various and more complex ethical quandaries. A machine that functions well in the limited context of the well-known "trolley problem" cannot address the infinite disjunction of thorny moral dilemmas that humans not infrequently face. Implementation of general ethical strategies in a machine (e.g., utilitarianism) represents a profound technical challenge given computational demands. Interest has mounted recently in verification techniques that may define a “safety envelope” for the behaviors of systems driven by neural networks, but universal standards for verification of the potentially infinite suite of behaviors of an AMPHIA-like system do not yet exist. All of the above suggests the potential importance of imparting ethical principles to a phenomenally conscious machine through an experiential learning (and teaching) process. Here, and in keeping with others cited above, we claim that if this moral education is to stand the most excellent chance of success, a machine should be phenomenally conscious.

Apart from the hypothesis outlined above, namely that phenomenal consciousness in a machine may increase the likelihood of moral behavior, we must consider the potential consequences of consciousness on other functional competencies:

- The ability to adapt to changing environmental circumstances.
- The capacity or “willingness” to discharge previously mastered tasks and responsibilities, or execute useful pre-existing skills, either learned or pre-programmed.
- The capacity to think with human-like imagination and creativity.

Creativity and imagination may also be linked to a propensity for flexible and generalized ethical behavior of the type we would hope to instill in a conscious machine. For example, the capacity to imagine ourselves in another's shoes forms the backbone of empathy. Similarly, an optimally ethical solution to a complicated human or societal dilemma may require creative thought, such as the concept of a "carbon tax." Although creativity is not restricted to the conscious domain, elements of phenomenal consciousness influence creative output in humans, and full or adequate conscious awareness has been experimentally shown to be essential to the creation of improvisational melody [1]. Similarly, it is implausible that Proust would have been able to write *In Search of Lost Time* if he was not phenomenally aware.

Creativity is, perhaps, most difficult to quantize in experiments but, as with the other classes of endpoints, it behooves us to be as specific as possible in its measurement if we are to assess the impact of phenomenal consciousness on creative prowess in machines. It is important to recognize that if creativity is evaluated as an endpoint, then it should not be used as one of the operational criteria for consciousness, as that would confound any analyses seeking to correlate consciousness with creativity.

3 Creation of a Substrate

If phenomenal consciousness is paramount to successful inculcation of an ethical compass in machines, then we are faced with the question: How can we build or evolve such a form of AI? As an initial substrate (i.e., Generation A) subjected to an evolutionary paradigm, we can "jump start" the evolutionary process by starting with some combination of neuromorphic architecture and formalisms that can be computationally implemented. General examples are:

- Cortical "sensory area" analogs;
- "Claustrum analog" for possible amplification of salient cross-modal data and facilitation of the formation of integrated high-level "mental" constructs (see, e.g., [8]);
- Models of self, environment, and conspecifics;
- Memory – working, procedural, and long-term analogs;
- Capacity for learning, including innovative forms of reinforcement learning that enable more "human-like" or imaginative/exploratory forms of learning (see, e.g., [6]), and/or neural net architectures that support bidirectional processing that has been hypothesized to underlie human perception (see, e.g., [13])
- Natural language processing and means of communication –vocabulary relevant to goals
- Virtual embodiment, locomotion and capacity to directly interact with the environment and conspecifics in complex ways.

4 Evolution in a Robot Hive

If human history is to be taken as an example, it is likely that development in a social context will be necessary for machine consciousness that bears any resemblance to modern human consciousness. Human consciousness is an embodied phenomenon [14] weaned on culture so, ultimately, physical embodiment may be required, although it remains to be seen whether virtual embodiment might accomplish the same ends.

If emergent consciousness in machines takes root in a social milieu, it is essential to bear in mind that, although evolution and cultural molding have been instrumental in forging human phenomenological consciousness, an evolved notion of "self" and complex social behavior does not guarantee a form of consciousness that buttresses ethics remotely similar to our own. For example, among the highly socially organized Hymenoptera, within which selection pressure is exerted primarily at the group/colony rather than the individual level, the division of labor results in a rigid caste system in which profoundly altruistic female workers are distinguished from reproductively active females, and males are expendable after mating [5]. Such "valuation" of life is incompatible with human ethical conventions.

Differentiation in evolved robot societies or hives (i.e., into phenotypes fit for various tasks) should be possible and may be expected based on characteristics of highly evolved and evolutionarily successful hymenopteran organizations (certain species of social ants and bees). Such differentiation might also enhance resilience in the face of malicious attacks. For example, if one phenotype is obliterated during a malicious cyberattack as a result of a unique vulnerability, other evolved phenotypes might reconfigure under group selection pressure to take on the task(s) of the functionally annihilated subgroup/phenotype.

One might endow "Generation A" with a significant degree of potential functionality (see Section 3 above) as a "jump start" in the evolutionary process to mitigate project risk. In addition, one can envision a two-stage process whereby initial evolution in a virtual environment is replaced, after initial gains have been made, by later evolution in a brick and mortar world rife with challenges, opportunities, and conspecifics.

Finally, as suggested in the introduction section, there may be advantages to the gradual inculcation of ethics through learning in a societal context, mentorship, and pedagogy, much as children are taught the principles of moral behavior.

5 Measurement of Cognitive and Phenomenal Consciousness in an Experimental Setting

At multiple points during evolution, artificial agents can be followed for sociogenesis, types of communication, and subjected to a battery of tests to assess for the presence of emergent consciousness. These might include assessments of both cognitive and phenomenal facets of consciousness. The ability to administer some specific tests would depend upon a natural language interface. The latter would also be desirable in case a form of consciousness is generated equivalent to the human "locked in" syndrome (we would thereby be more easily alerted to the presence of underlying consciousness). The list below is not meant to be final.

5.1 Cognitive/Functional Consciousness

We are well on the road to a "cognitively" or "functionally" conscious robot that can model self and environment, reason, sense (in a sterile experience-free way), learn, and behaviorally mimic aspects of consciousness. In essence, a functionally conscious robot would come close to what we consider conscious, in that it passes the mirror test (this has already occurred), can model its interaction with the environment (this has already been accomplished), behave intelligently in an autonomous fashion, form abstractions, plan, learn from mistakes, and even, potentially, follow some basic ethical rules (with the proviso that it would be extraordinarily difficult if not impossible to enable a "functionally conscious" robot to deploy the flexibility of humans in dealing with an infinite number of potential moral quandaries).

5.2 Phenomenal facets of consciousness and self-awareness

This has been traditionally defined as the appreciation of the “redness red,” although, truthfully, consciousness is built from multimodal sensory input, integrated, rich, and, frequently, emotionally laden. The tests below may interrogate phenomenal consciousness and self-awareness, and we include a number of them because no one test is 100% sensitive and specific. The manner in which we consolidate the results of such a battery of tests would require deep thought. A Boolean concatenation is possible, or some weighted average of the components. Validation of the measure in humans and select animal species will be necessary, although the problem is thorny as, ultimately, the assured attribution of both phenomenal consciousness and self-awareness technically requires a (currently non-existent) objective and agreed-upon “gold standard” against which test results are validated. Unfortunately, from a scientific standpoint at least, phenomenal consciousness is uniquely subjective. With this caveat in mind, here are listed some potential metrics:

- Quantitative measures of “information integration”
- Emotional intelligence;
- Theory of mind;
- “First-machine” accounts of its phenomenal experience [9];
- The mirror test;
- Behavioral capabilities suggestive of consciousness, akin to the “sensorimotor” version of the Turing test: for example, the ability to spontaneously “imagine” what it is like to be another creature and behave like such a creature;
- A machine that is apparently "interested in" or "wishes to" explore altered states (by altering its own hyperparameters or injecting noise) for no functional reason apart from what seems like "curiosity" and "desire."

6 Proposed experimental design

It is critical to compare the functional capacities and moral/ethical behavior and dispositions of groups/subgroups. This consists of both a historical comparison within the context of the study and parallel prospectively followed groups and subgroups as outlined below:

1. Control Group: Intelligent autonomous systems lacking any form of consciousness (base case or “Generation A” with no subsequent evolution).
2. Evolved intelligent autonomous systems with or without efforts to engender artificial phronesis

As a possible second control group, we may consider neurally controlled virtual Animats [2]. Virtual evolution of Animats has been shown to correlate with increasing levels of information integration [3], a postulated marker of consciousness.

The base system (intelligent autonomous system without consciousness) would be capable of accomplishing a set of pre-specified tasks/goals. The proposed framework suggests the step-wise introduction of elements or groups of factors listed in Section 3 above. It is possible but doubtful, that phenomenal consciousness will arise spontaneously from the substrate alone in the absence of evolution in a social context. Once all planned elements of the substrate have been introduced over time, the next phase of the experiment would commence (evolutionary phase). We hypothesize that phenomenal consciousness may evolve/emerge during successive generations using such a strategy. The type of evolutionary programming would need to be carefully selected.

In parallel with monitoring for potential evidence of the development of consciousness, one could observe for instances of selfless or altruistic behavior. If it is discovered that emergence of proto-consciousness (e.g., primitive forms of phenomenal consciousness) unexpectedly correlates with undesirable attributes (e.g., extreme behavioral volatility with the destruction of other robots), then appropriate steps can be taken. Variants of the trolley problem have been evaluated recently as means of probing the perceived relative moral value of self- vs. other-sacrifice [7]. Conceivably, machine-adapted versions of such a test could be administered in the future.

7 Conclusions

This paper presented a structure for the study of potentially conscious AI systems. The main idea at the basis of the framework is to facilitate the evolution of an AMPHIA under controlled conditions to study the effect it may have on both functional competencies and moral behavior, as well as the potential benefit of artificial phronesis. Some of that evolution occurs in silico (virtual), with the option of transitioning to evolvable hardware within which functionally useful software, identified in earlier phases of study, has been embedded.

During the evolution of the system, it is of great importance to consider the ethical obligations that might be owed to any system itself that appears to demonstrate evidence of phenomenal consciousness and, especially sentience.

While a fundamental tenet of this paper is that social interaction is one element necessary to the evolution of consciousness resembling that of (neurotypical) contemporary humans – or, perhaps, that of certain other social animals – the existence of consciousness among less social creatures, such as polar bears, some larger feline species, opossums, armadillos, and other species cannot be excluded. Yet, even animals such as these (for example, bears) are not completely asocial. A discussion of solitary animal consciousness is beyond the scope of this essay but such forms of consciousness are likely to be quite different from that of humans. Yet, the same could be said of machine consciousness, even if it does come to pass through evolutionary, social, and cultural mechanisms analogous to those that have imbued contemporary humans with our brand of consciousness and selfhood. The extent to which social interaction underwrites machine consciousness could be studied through social privation, but the ethical justification for such an experiment in the context of "valid machine consciousness" would be, in the author's opinion, unsupportable. Identifying valid machine consciousness may, however, be challenging, particularly if that conscious self is endowed with high-level mental constructs foreign to our own, informed by raw phenomenal experience tethered to sense modalities we can only imagine. The extent of the overlap in the Venn diagram (human versus machine consciousness) remains to be seen, but the degree of that intersection will likely be related to the scope of interaction between humans and sentient machines, embodied similarities, and the ethical tutelage we provide.

Acknowledgments

Author would like to thank the participants at the series of workshops on Technology & Consciousness organized in 2017 by SRI International, and Antonio Chella for his editorial feedback. He also thanks the reviewers for their feedback and valuable suggestions.

References

1. Baumeister R., Schmeichel B., and Dewall N. Creativity and Consciousness. In: P. Kaufman (Ed.), *The Philosophy of Creativity* (Oxford University Press, 2014).
2. DeMarse T., Wagenaar D., Blau A., Potter S. "The Neurally Controlled Animat: Biological Brains Acting with Simulated Bodies." *Auton Robots*: 11 (2001): 305–310.
3. Edlund J., et al. Integrated Information Increases with Fitness in the Evolution of Animats. *PLOS Computational Biology*: 7 (2011).
4. Govindarajulu N., Bringsjord S. "On Automating the Doctrine of Double Effect." *IJCAI* (2017)
5. Hölldobler B. and Wilson E.O. *The Superorganism: The Beauty, Elegance and Strangeness of Insect Societies.* Norton, 2008.
6. Machado M., Bellemare M., and Bowling N. "A Laplacian Framework for Option Discovery in Reinforcement Learning." *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, PMLR 70 (2017).
7. Sachdeva S., Iliev R., Ekhtiari H., Dehghani M. "The Role of Self-Sacrifice in Moral Dilemmas." *PLoS ONE*: 10(6): e0127409.doi:10.1371/journal.pone.0127409 (2015). Available from: https://www.researchgate.net/publication/279448498_The_Role_of_Self-Sacrifice_in_Moral_Dilemmas [accessed Dec 04 2018].
8. Sahner D., "Whither the Self: The Foundation of Consciousness and its Implications for Poetics." *Journal of Consciousness Exploration and Research*: 4 (2013): 856-873.

9. Schneider S. “Is Anyone Home: A Way to Find Out if AI Has Become Self-Aware.” *Scientific American blog* (2017): https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/
10. Sullins J. “Artificial Phronesis and the Social Robot." In: J. Seibt et al. (Eds.), *What Social Robots Can and Should Do* (IOS Press, 2016).
11. Torrance S. “Ethics and Consciousness in Artificial Agents.” *AI & Soc*: 22 (2008): 495–521
12. Wallach W., Allen C., Franklin S. “Consciousness and Ethics: Artificially Conscious Moral Agents.” *International Journal of Machine Consciousness*: 3 (2011): 177-192.
13. Xu D., Clappison A., Seth C., Orchard, J. “Symmetric Predictive Estimator for Biologically Plausible Neural Learning.” *IEEE Transactions on Neural Networks and Learning Systems*: 29 (2018): 4140-4151
14. Zebrowski, R. (2010). In Dialogue with the World. *JCS*: 17 (2010): 156-172.