Integrative Biological Simulation, Neuropsychology, and AI Safety

Gopal P. Sarma[1*], **Adam Safron**[2], **and Nick J. Hay**[3†]
School of Medicine, Emory University, Atlanta, GA USA
Department of Psychology, Northwestern University, Evanston IL USA
Vicarious AI, San Francisco, CA USA

Abstract

We propose a biologically-inspired research agenda with parallel tracks aimed at AI and AI safety. The bottom-up component consists of building a sequence of biophysically realistic simulations of simple organisms such as the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and the zebrafish *Danio rerio* to serve as platforms for research into AI algorithms and system architectures. The top-down component consists of an approach to value alignment that grounds AI goal structures in neuropsychology. Our belief is that parallel pursuit of these tracks will inform the development of value-aligned AI systems that have been inspired by embodied organisms with sensorimotor integration. An important set of side benefits is that the research trajectories we describe here are grounded in long-standing intellectual traditions within existing research communities and funding structures. In addition, these research programs overlap with significant contemporary themes in the biological and psychological sciences such as data/model integration and reproducibility.

Introduction

Bostrom's orthogonality thesis states, under certain weak assumptions, that the intelligence of an agent and its goal structure are independent variables (Bostrom 2014). The orthogonality thesis is a useful conceptual tool for correcting for anthropomorphic bias, i.e. the assumption that an arbitrary agent will behave in a manner similar to human beings. Particularly in a climate of fear and uncertainty about future AI systems, the orthogonality thesis can be a helpful framing to encourage reasoning more clearly about the risks of advanced AI systems and to dislodge concerns arising from science fiction movies and sloppy journalism. However, from an engineering standpoint, it is worth considering that the design of safe, superintelligent AI systems may *benefit* from examining system architectures in which the intelligent substrate and fundamental goal structure of an agent have been intentionally coupled. We have used the phrase *anthropomorphic design* to refer to approaches in which AI systems are built to possess commonalities with human neuropsychology (Sarma and Hay 2017; Sarma, Hay, and Safron 2018; Sotala 2016).

We have previously argued that research in affective neuroscience and related disciplines aimed at grounding human values in neuropsychology may provide important conceptual foundations for understanding value alignment in AI systems. There are several practical benefits that might emerge from such a research program. Having more detailed prior information about human values may allow a sophisticated AI system to learn from fewer examples or enable practical implementations of AI safety techniques that would otherwise be computationally intractable (Sarma and Hay 2017; Sarma, Hay, and Safron 2018).

In this brief position paper, we propose a complementary agenda consisting of a bottom-up approach to understanding AI architectures. This program is aimed at building realistic biophysical simulations of simple nervous systems which incorporate biomechanics in a simulated environment. We believe that the parallel pursuit of these goals will not only lead to fundamental advances in AI algorithms, but also architectural insights into ensuring value alignment. The contributions of this manuscript are two-fold: *(i)* We introduce a set of research objectives in neuroscience that are well positioned to give rise to significant advances in AI and which have received little attention by the AI safety community. *(ii)* We suggest two existing approaches to AI safety that may integrate with this research paradigm: a neuropsychology-based approach to value alignment and *test suites* for agent-based AI systems in simulated environments.

Integrative Biological Simulation

Claim 1: Simple organisms show complex behavior that continues to be difficult for modern ML systems. Neuronal simulations in virtual environments will allow these biological architectures to be used for AI research.

Integrative biological simulations refer to computational platforms in which diverse, process-specific models, often operating at different scales, are combined into a global, composite model (Sarma and Faundez 2017). Examples include OpenWorm, an internationally coordinated open-science project working towards a realistic biophysical

[*] Email: gopal.sarma@emory.edu

[†] The views expressed herein are those of the author and do not necessarily reflect the views of Vicarious AI.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27321v1 | CC BY 4.0 Open Access | rec: 6 Nov 2018, publ: 6 Nov 2018

simulation of the nematode *Caenorhabditis elegans*, Neurokernel, a project with some parallels to OpenWorm aimed at simulating *Drosophila melanogaster*, Virtual Lamprey, a computational platform for understanding vision and locomotion in the lamprey, BlueBrain, an effort to build a detailed model of the rat cortical micro-column, and the Human Brain Project (HBP), an ambitious successor to BlueBrain which aims to extend this platform to an entire human cerebral cortex (Sarma et al. 2018; Givon and Lazar 2016; Sarvestani et al. 2013; Markram et al. 2015; Amunts et al. 2016). Such platforms may serve as points of integration for data and computational models. The result is a shared structure that can be used by an entire community of researchers to test novel hypotheses, create a tighter feedback loop between experimental and theoretical research, and ensure the reproducibility and robustness of the underlying research output.

In the AI community, awareness of these research programs has primarily been informed by the efforts of BlueBrain and HBP to simulate large regions of mammalian cortical tissue. We are sympathetic in many ways to the aims of these projects. However, we believe that an under-appreciated set of approaches complementing their work consists of using analogous software infrastructure to develop simulations of organisms far below the complexity of mammals or vertebrates. *C. elegans*, with only 302 neurons, shows simple behavior of learning and memory. *Drosophila melanogaster*, despite only having 10^5 neurons and no comparable structure to a cerebral cortex, has sophisticated spatial navigation abilities easily rivaling the best autonomous vehicles with a minuscule fraction of the power consumption. The zebrafish *Danio rerio* has on the order of 10^7 neurons and has been a model system in neuroscience for several decades. Moreover, recent efforts to perform whole-brain functional imaging in the larval zebrafish may make this a particularly attractive target for future integrative simulation platforms (Ahrens et al. 2013). Although much of this research has been motivated by neuroscientific aims and connections to the study of human disease processes, the implications for AI research are significant. Well-engineered software platforms which allow for rapid iteration on existing architectures without the constraints of biological realism will allow AI researchers to test novel hypotheses in embodied organisms in simulated environments. Real-time visualization of nervous system activity will allow for a deeper understanding of how AI algorithms such as backpropagation, belief propagation, or reinforcement learning may approximate what is observed in nature.

Coupling nervous system activity to drive a simulated body is a tractable approach with organisms such as *C. elegans* and *Drosophila*. In OpenWorm, for example, the Boyle-Cohen model of neuromuscular coupling allows for the output of connectome dynamics to drive the activation of body wall muscles and a simulated body (Gleeson et al. 2018; Palyanov, Khayrulin, and Larson 2018). Similar models are likely achievable with *Drosophila* as well. Indeed, the Neurorobotics Platform of HBP is working towards a general platform for interfacing realistic neural network simulations with robotic bodies (Falotico et al. 2017). The incorporation of biomechanics into these simulations can be justified on biological grounds. For instance, understanding the effects of anti-psychotic or anti-epileptic medications is simplified if researchers can observe changes in behavioral patterns, rather than having to interpret high-dimensional data streams of neuronal activity. However, there are reasons to think that sensorimotor integration may be particularly valuable from a purely AI perspective.

As others have argued, despite the significant advances arising from the use of deep representations in neural networks, current AI systems continue to lack many of the qualities of fluid intelligence observed in human beings, particularly in the ability to learn concepts from a relatively small number of examples. One hypothesis is that human concepts are grounded in rich sensorimotor experience. Despite significant work in transfer learning and domain adaptation, modern systems are largely restricted in their domain of application. The lack of behavior-based concept representation may be a limiting factor in current state-of-the-art systems (Hay et al. 2018; Krichmar 2018; Falotico et al. 2017). Simulations of simple, embodied organisms with realistic virtual environments may provide platforms for AI research aimed at understanding the interplay between concept representation and embodiment. Moreover, used in a modular or hierarchical fashion, contemporary techniques such a deep learning may prove to be powerful components of future iterations of these platforms.

Neuropsychology and Value Alignment

Claim 2: Value-alignment research may benefit from insights in neuropsychology and comparative neuroanatomy.

We have argued previously for an approach to value alignment which grounds an understanding of human values in neuropsychology (Sarma and Hay 2017; Sarma, Hay, and Safron 2018). In this section, we reproduce the broad outlines of this framework before discussing how these parallel research tracks may come to intersect. Our approach is loosely based on research in affective neuroscience, which aims to categorize emotional universals in the mammalian kingdom and correlate them with an underlying neurological substrate (Panksepp 1998). Other related and possibly relevant fields of research include contemplative neuroscience, neuropsychoanalysis, biological anthropology, and comparative neuroanatomy, to name just a few.

It is possible that values and motivations are grounded in emotions for human beings. If our emotional substrate is shared with other mammals, or even more broadly with other vertebrates and animals, it suggests that our value systems can be decomposed in ways that inform neuroscience-based

AI architectures. For example, one possible decomposition of human values is the following:

1. **Internal reward systems shared by all mammals**: In the taxonomy of affective neuroscience, these include play, panic/grief, fear, rage, seeking, lust, and care. This may also include curiosity and the acquisition of skills.
2. **Internal reward systems unique to human beings**: For example, uniquely human social behaviors such as family membership, group affiliation, and story telling.
3. **Products of human deliberation/cognition on our values:** The many complex features of value systems produced by several millennia of human social and cultural evolution; likely mediated by cultural inheritance.

Decompositions such as this one might allow AI systems to begin with a more nuanced understanding of human values that is then refined over time through observation, hypothesis generation, and human interaction. For an agent that is actively interacting with the world during the learning process, a more informative prior may allow a system to learn from fewer examples, directly translating into a reduced risk of adverse outcomes. Likewise, consider that our values and culture are instilled in children by selective exposure to carefully chosen environments. A neuropsychological understanding of human values may allow us to make similarly strategic choices for AI systems in order to minimize the time required to achieve strong guarantees of value alignment (Evans, Stuhlmüller, and Goodman 2016; Christiano et al. 2017). Moreover, systems with human-inspired architectures may lead to natural avenues for addressing issues of transparency and intelligibility of AI decision making (Wortham, Theodorou, and Bryson 2017; Wachter, Mittelstadt, and Floridi 2006).

Synthesis

Claim 3: Significant synergy may be achieved by coupling the two research programs described above.

Thus far, we have discussed organisms which lie very far apart on the evolutionary tree. *C. elegans* and *Drosophila* possess only 10^2 - 10^5 neurons and the zebrafish *Danio rerio* roughly 10^7 neurons, whereas mammalian brains range from 10^8 neurons in the brown rat to 10^{10} neurons in human neocortex. However, by the time we reach *Drosophila* we are already confronting a brain with many high-level architectural features which higher animals share, such as two lobes and distinct functional processing regions. Moreover, insects share many neurochemical motivation systems with vertebrates and even higher mammals (Panksepp 1998). Proceeding up the evolutionary tree a little further, sophisticated brain centers involved in motor coordination, such as the basal ganglia, are known to be conserved across vertebrates (including zebrafish), and may have homologous structures in arthropods (Grillner and Robertson 2016). In other words, viewed as platforms for research into value-aligned AI systems, there may be clues even from invertebrates and simple vertebrates for how the insights from top-down, neuropsychology-based approaches may be used to design AI systems that possess far greater levels of transparency, intelligibility, and goal structure stability than we see in nature or in our current AI technologies. Moreover, BlueBrain/HBP are already tackling the substantially more difficult challenge of simulating mammalian brains. Our proposed agenda is in no way in opposition to their work. Indeed, the success of BlueBrain/HBP will only complement insights that arise from approaches oriented towards simulation of simple organisms.

Another point of intersection between integrative biological simulation and current research in AI safety is to extend the concept of test suites for RL agents to the virtual environments of simulated organisms (Leike et al. 2017). Problems such as safe interruptibility, avoiding negative side effects, reward gaming, distributional shift, and others should be adaptable to virtual biological organisms. For example, to what degree do we see variation in susceptibility to reward hacking (i.e. addictive behaviors) in the animal kingdom? Lifting biological constraints, can we augment simulated architectures with modules to reduce the risk of such behavior?

Discussion and Future Directions

We encourage interested researchers to consult the publications of the respective simulation projects to find concrete points of entry. For those attracted to expanding the repertoire of simple organisms that have such platforms, there are many commonalities in the necessary software infrastructure, with tools such as NEURON for simulating Hodgkin-Huxley type models, BluePyOpt for extracting kinetic parameters for experimental data, and NetPyNE/Bionet for specifying network models (Hines and Carnevale 1997; Van Geit et al. 2016; Gratiy et al. 2018). Aside from the connectome, an area where there are relevant differences between these organisms is in the gene expression of ion channels. Efforts such as ChannelPedia, NeuroMLDB, ModelDB, and Open Source Brain, all share the goal of enabling storage and re-use of neuroscience data and models (Ranjan et al. 2011; Gleeson et al. 2012). Expanding the scope of these resources to include ion channel data and models for a variety of species would be a key enabler of this research agenda. We suspect that there is literature on comparative neuroanatomy that will give us insights into promising directions to pursue on the lower part of the evolutionary tree.

With regards to the top-down approach to value-alignment, as we emphasized in our previous manuscript, a key obstacle is the widespread concern of reproducibility issues in the biological and psychological literature (Sarma, Hay, and Safron 2018). Therefore, we are of the conviction that the most immediate next step is to create a community-driven replication effort aimed at developing a more robust body of knowledge with which to base future research. To that end, we have created a project using the Open Science Framework where we are currently collecting suggestions for candidate studies which would be of high value to either directly replicate or validate through some other means.[1] We are particularly interested in using iterated expert elicitation methods such as RAND Corporation's Delphi protocol to encourage consensus building among researchers.

Finally, regarding the development of test suites for simulated organisms, we have no illusions as to the difficulty of the challenge. Understanding how to translate the highly simplified models of current AI safety frameworks to the complex neural networks of real organisms in realistic physical environments will be a substantial undertaking. However, we believe that such a synthesis is both necessary and desirable, as it may provide insight into hybrid approaches which take advantage of both modern AI and simulated biology to build sophisticated value-aligned systems.

Acknowledgments

We would like to thank Owain Evans and Tom Everitt for insightful discussions and feedback on the manuscript.

References

Ahrens, M. B.; Orger, M. B.; Robson, D. N.; et al. 2013. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods* 10(5):413.

Amunts, K.; Ebell, C.; Muller, J.; et al. 2016. The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain. *Neuron* 92(3):574–581.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *NIPS*, 4299–4307.

Evans, O.; Stuhlmüller, A.; and Goodman, N. D. 2016. Learning the preferences of ignorant, inconsistent agents. In *AAAI*, 323–329.

Falotico, E.; Vannucci, L.; Ambrosano, A.; et al. 2017. Connecting Artificial Brains to Robots in a Comprehensive Simulation Framework: The Neurorobotics Platform. *Frontiers in Neurorobotics* 11:2.

Givon, L. E., and Lazar, A. A. 2016. Neurokernel: an open source platform for emulating the fruit fly brain. *PLOS ONE* 11(1):e0146581.

Gleeson, P.; Piasini, E.; Crook, S.; et al. 2012. The Open Source Brain Initiative: Enabling Collaborative Modelling in Computational Neuroscience. *BMC Neuroscience* 13(1):O7.

Gleeson, P.; Lung, D.; Grosu, R.; et al. 2018. c302: A multiscale framework for modelling the nervous system of Caenorhabditis elegans. *Phil. Trans. R. Soc. B* 373(1758):20170379.

Gratiy, S. L.; Billeh, Y. N.; Dai, K.; et al. 2018. BioNet: A Python interface to NEURON for modeling large-scale networks. *PLOS ONE* 13(8):e0201630.

Grillner, S., and Robertson, B. 2016. The Basal Ganglia over 500 Million Years. *Current Biology* 26(20):R1088–R1100.

Hay, N.; Stark, M.; Schlegel, A.; et al. 2018. Behavior is Everything–Towards Representing Concepts with Sensorimotor Contingencies. In *AAAI*.

Hines, M. L., and Carnevale, N. T. 1997. The NEURON Simulation Environment. *Neural Computation* 9(6):1179–1209.

Krichmar, J. L. 2018. Neurorobotics—A Thriving Community and a Promising Pathway Toward Intelligent Cognitive Robots. *Front. Neurorobot.* 12.

Leike, J.; Martic, M.; Krakovna, V.; et al. 2017. AI Safety Gridworlds. *arXiv preprint arXiv:1711.09883*.

Markram, H.; Muller, E.; Ramaswamy, S.; et al. 2015. Reconstruction and Simulation of Neocortical Microcircuitry. *Cell* 163(2):456–492.

Palyanov, A.; Khayrulin, S.; and Larson, S. D. 2018. Three-dimensional simulation of the Caenorhabditis elegans body and muscle cells in liquid and gel environments for behavioural analysis. *Phil. Trans. R. Soc. B* 373(1758):20170376.

Panksepp, J. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.

Ranjan, R.; Khazen, G.; Gambazzi, L.; et al. 2011. Channelpedia: an integrative and interactive database for ion channels. *Front. Neuroinform.* 5:36.

Sarma, G. P., and Faundez, V. 2017. Integrative biological simulation praxis: Considerations from physics, philosophy, and data/model curation practices. *Cellular Logistics* 7(4):e1392400.

Sarma, G. P., and Hay, N. J. 2017. Mammalian Value Systems. *Informatica* 41(4).

Sarma, G. P.; Lee, C. W.; Portegys, T.; et al. 2018. OpenWorm: Overview and recent advances in integrative biological simulation of Caenorhabditis elegans. *Phil. Trans. R. Soc. B* 373(1758):20170382.

Sarma, G. P.; Hay, N. J.; and Safron, A. 2018. AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values. In *International Conference on Computer Safety, Reliability, and Security*, 507–512. Springer.

Sarvestani, I.; Kozlov, A.; Harischandra, N.; et al. 2013. A computational model of visually guided locomotion in lamprey. *Biological Cybernetics* 107(5):497–512.

Sotala, K. 2016. Defining Human Values for Value Learners. In *AAAI Workshop: AI, Ethics, and Society*.

Van Geit, W.; Gevaert, M.; Chindemi, G.; et al. 2016. BluePyOpt: Leveraging Open Source Software and Cloud Infrastructure to Optimise Model Parameters in Neuroscience. *Front. Neuroinform.* 10:17.

Wachter, S.; Mittelstadt, B.; and Floridi, L. 2006. Transparent, Explainable, and Accountable AI for Robotics. *Science Robotics* 2.

Wortham, R. H.; Theodorou, A.; and Bryson, J. J. 2017. Improving robot transparency: real-time visualisation of robot AI substantially improves understanding in naive observers. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 1424–1431.

[1] https://tinyurl.com/AI-reproducibility