Integrative Biological Simulation, Neuropsychology, and AI Safety

Gopal Sarma[1,2] PhD, Adam Safron[3] PhD, and Nick J. Hay[4] PhD

1. Emory University School of Medicine, Atlanta, GA USA
2. The OpenWorm Foundation, Boston, MA USA
3. Northwestern University, Evanstan, IL USA
4. Vicarious AI, San Francisco, CA USA

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

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

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

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

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

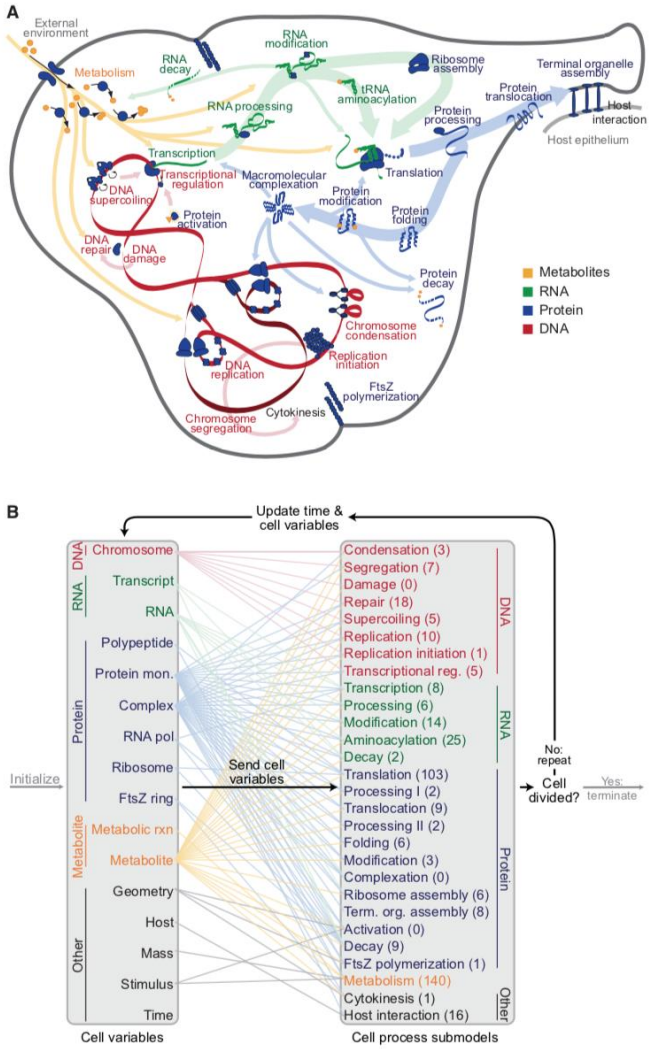
- **Claim 3:** *Significant synergy may be achieved by coupling the two research programs described above.*

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

Integrative Biological Simulations

Integrative Biological Simulations

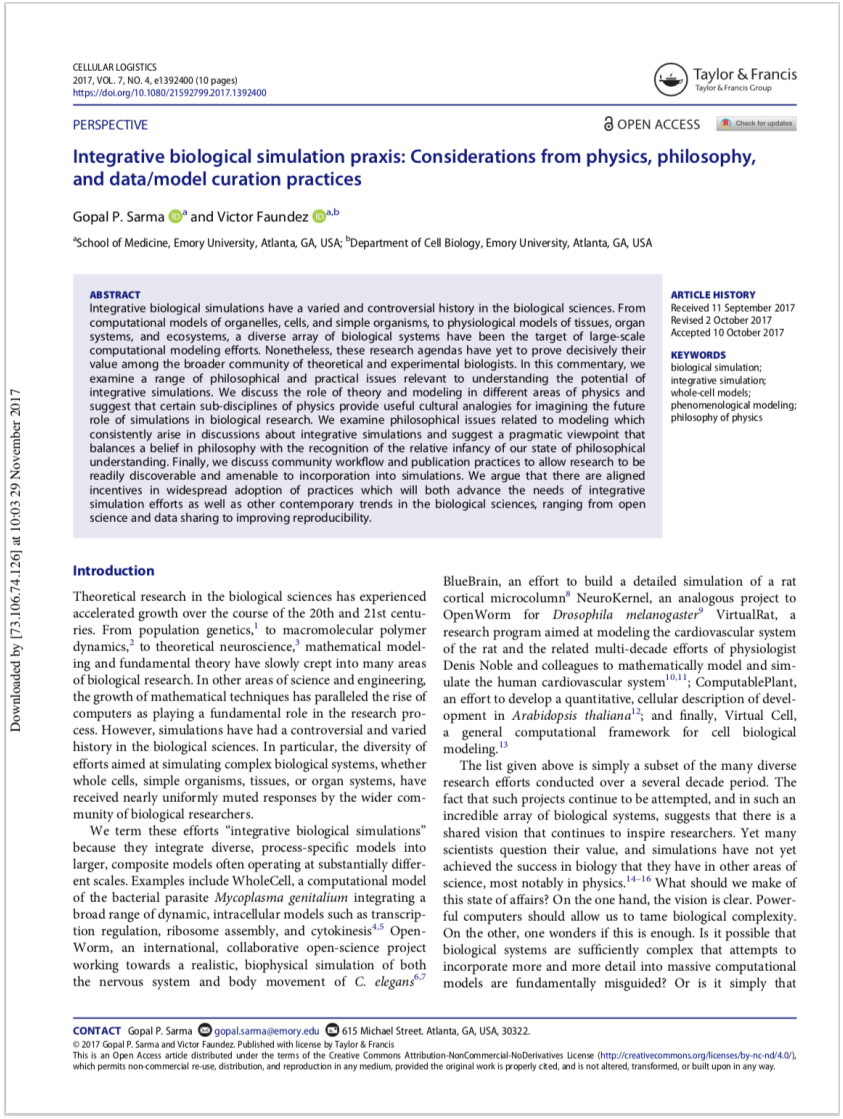
- Computational platforms in which diverse, process-specific models, often operating at different scales, are combined into a global, composite model



Karr, Jonathan R., et al. "A whole-cell computational model predicts phenotype from genotype." *Cell* 150.2 (2012): 389-401.

Integrative Biological Simulations

- Computational platforms in which diverse, process-specific models, often operating at different scales, are combined into a global, composite model
- **Why Develop These Platforms for the Life Sciences?** Complexity of biological systems; complexity of the corresponding research community; reproducibility and research quality



CELLULAR LOGISTICS
2017, VOL. 7, NO. 4, e1392400 (10 pages)
https://doi.org/10.1080/21592799.2017.1392400

Taylor & Francis
Taylor & Francis Group

PERSPECTIVE

OPEN ACCESS

Check for updates

Integrative biological simulation praxis: Considerations from physics, philosophy, and data/model curation practices

Gopal P. Sarma [a] and Victor Faundez [a,b]

[a]School of Medicine, Emory University, Atlanta, GA, USA; [b]Department of Cell Biology, Emory University, Atlanta, GA, USA

ABSTRACT

Integrative biological simulations have a varied and controversial history in the biological sciences. From computational models of organelles, cells, and simple organisms, to physiological models of tissues, organ systems, and ecosystems, a diverse array of biological systems have been the target of large-scale computational modeling efforts. Nonetheless, these research agendas have yet to prove decisively their value among the broader community of theoretical and experimental biologists. In this commentary, we examine a range of philosophical and practical issues relevant to understanding the potential of integrative simulations. We discuss the role of theory and modeling in different areas of physics and suggest that certain sub-disciplines of physics provide useful cultural analogies for imagining the future role of simulations in biological research. We examine philosophical issues related to modeling which consistently arise in discussions about integrative simulations and suggest a pragmatic viewpoint that balances a belief in philosophy with the recognition of the relative infancy of our state of philosophical understanding. Finally, we discuss community workflow and publication practices to allow research to be readily discoverable and amenable to incorporation into simulations. We argue that there are aligned incentives in widespread adoption of practices which will both advance the needs of integrative simulation efforts as well as other contemporary trends in the biological sciences, ranging from open science and data sharing to improving reproducibility.

ARTICLE HISTORY
Received 11 September 2017
Revised 2 October 2017
Accepted 10 October 2017

KEYWORDS
biological simulation; integrative simulation; whole-cell models; phenomenological modeling; philosophy of physics

Introduction

Theoretical research in the biological sciences has experienced accelerated growth over the course of the 20th and 21st centuries. From population genetics,[1] to macromolecular polymer dynamics,[2] to theoretical neuroscience,[3] mathematical modeling and fundamental theory have slowly crept into many areas of biological research. In other areas of science and engineering, the growth of mathematical techniques has paralleled the rise of computers as playing a fundamental role in the research process. However, simulations have had a controversial and varied history in the biological sciences. In particular, the diversity of efforts aimed at simulating complex biological systems, whether whole cells, simple organisms, tissues, or organ systems, have received nearly uniformly muted responses by the wider community of biological researchers.

We term these efforts "integrative biological simulations" because they integrate diverse, process-specific models into larger, composite models often operating at substantially different scales. Examples include WholeCell, a computational model of the bacterial parasite *Mycoplasma genitalium* integrating a broad range of dynamic, intracellular models such as transcription regulation, ribosome assembly, and cytokinesis[4,5] OpenWorm, an international, collaborative open-science project working towards a realistic, biophysical simulation of both the nervous system and body movement of *C. elegans*[6,7] BlueBrain, an effort to build a detailed simulation of a rat cortical microcolumn[8] NeuroKernel, an analogous project to OpenWorm for *Drosophila melanogaster*[9] VirtualRat, a research program aimed at modeling the cardiovascular system of the rat and the related multi-decade efforts of physiologist Denis Noble and colleagues to mathematically model and simulate the human cardiovascular system[10,11]; ComputablePlant, an effort to develop a quantitative, cellular description of development in *Arabidopsis thaliana*[12]; and finally, Virtual Cell, a general computational framework for cell biological modeling.[13]

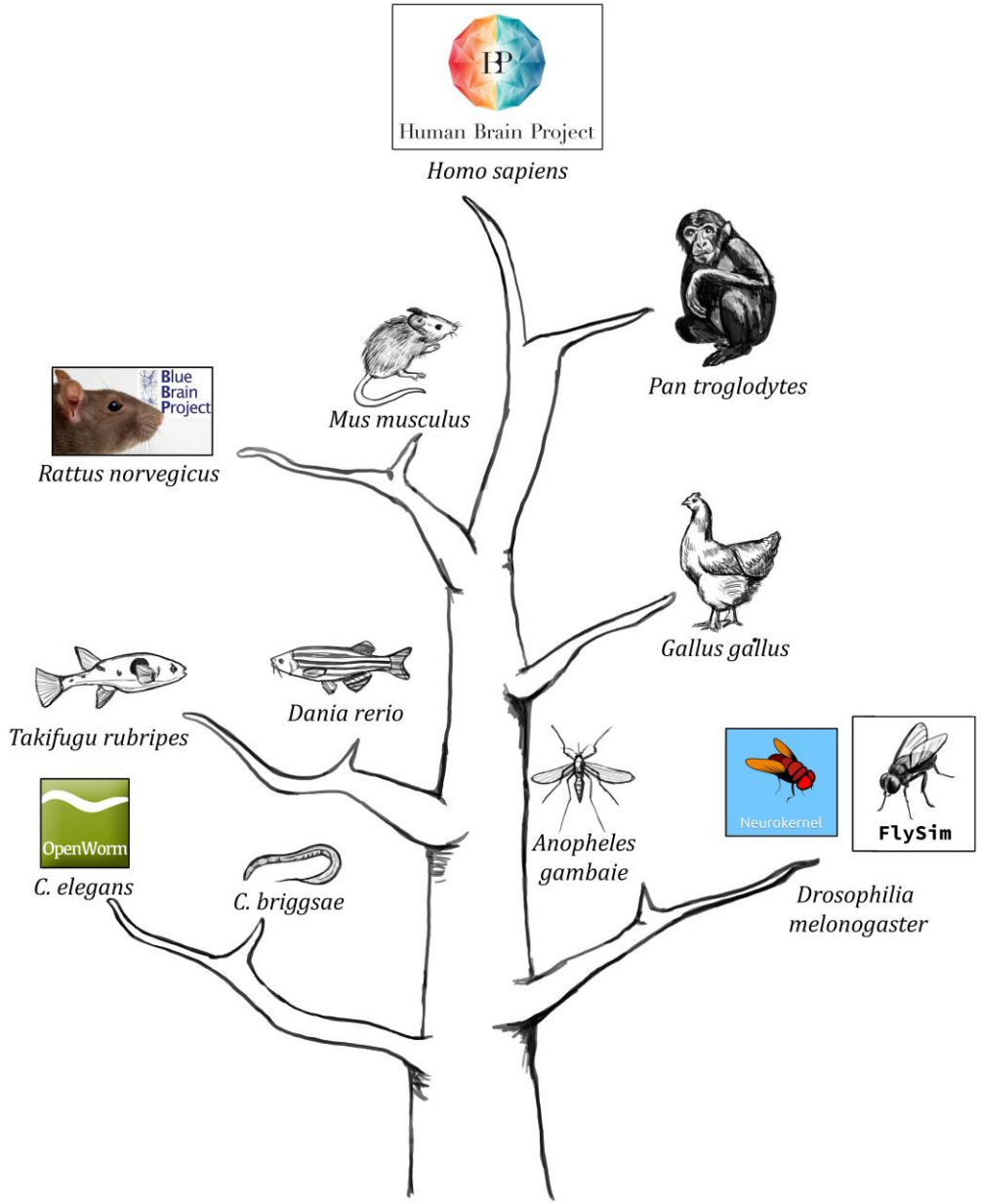
The list given above is simply a subset of the many diverse research efforts conducted over a several decade period. The fact that such projects continue to be attempted, and in such an incredible array of biological systems, suggests that there is a shared vision that continues to inspire researchers. Yet many scientists question their value, and simulations have not yet achieved the success in biology that they have in other areas of science, most notably in physics.[14–16] What should we make of this state of affairs? On the one hand, the vision is clear. Powerful computers should allow us to tame biological complexity. On the other, one wonders if this is enough. Is it possible that biological systems are sufficiently complex that attempts to incorporate more and more detail into massive computational models are fundamentally misguided? Or is it simply that

CONTACT Gopal P. Sarma gopal.sarma@emory.edu 615 Michael Street. Atlanta, GA, USA, 30322.

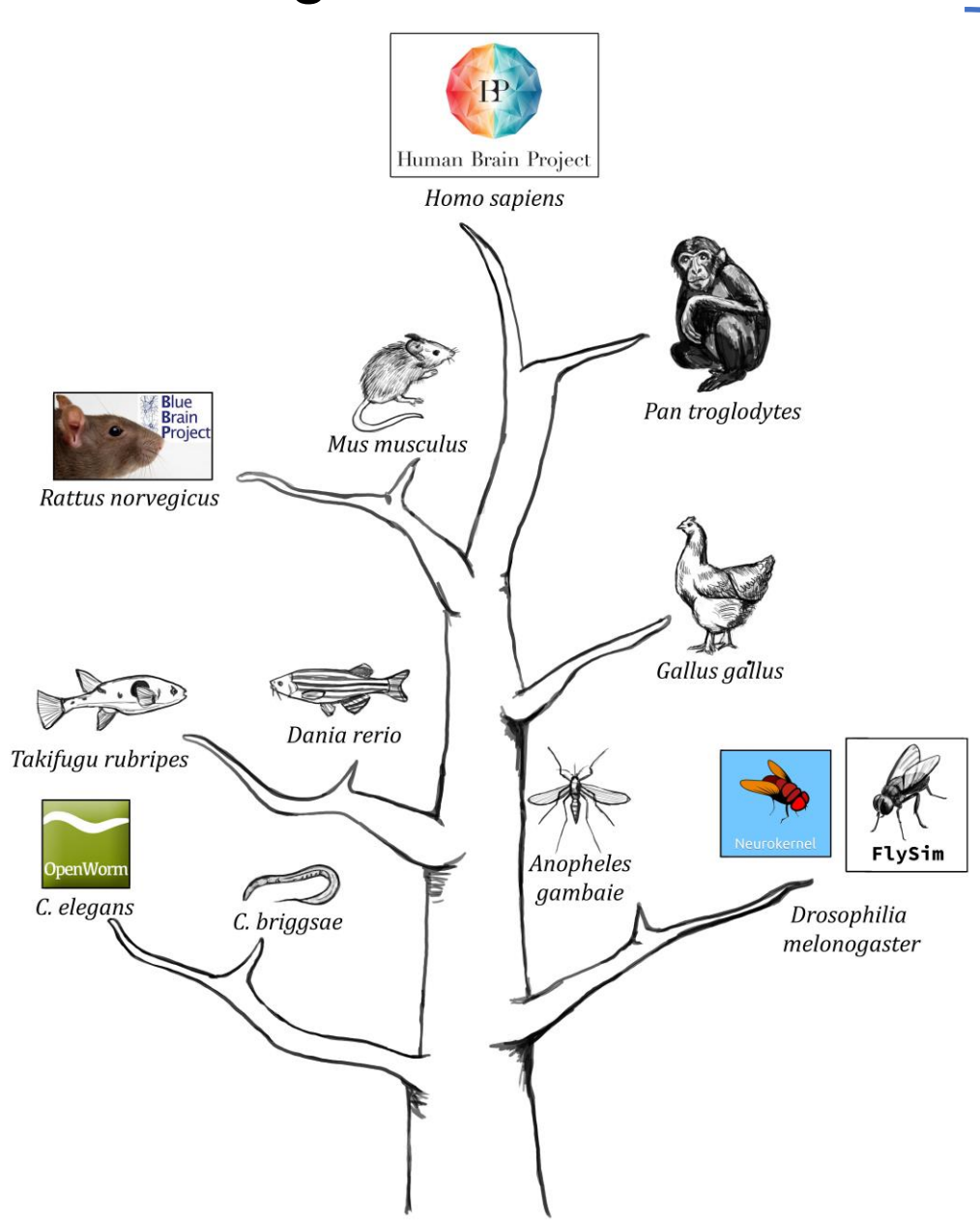
Sarma, Gopal P., and Victor Faundez. "Integrative biological simulation praxis: Considerations from physics, philosophy, and data/model curation practices." *Cellular Logistics* 7.4 (2017): e1392400.

Integrative Biological Simulation of Realistic Nervous Systems

Integrative Biological Simulation of Realistic Nervous Systems



Integrative Biological Simulation of Realistic Nervous Systems



NEURON

BluePyOpt

NetPyNE

Bionet

Geppetto

ChannelPedia

NeuroMLDB

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

Neuropsychology and AI Safety

Neuropsychology and AI Safety

- View human values from the perspective of neuropsychological foundations

Mammalian Value Systems

Gopal P. Sarma
School of Medicine, Emory University, Atlanta, GA USA
E-mail: gopal.sarma@emory.edu

Nick J. Hay
Vicarious FPC, San Francisco, CA USA
E-mail: nnickhay@gmail.com

Keywords: friendly AI, value alignment, human values, biologically inspired AI, human-mimetic AI

Received: June 24, 2013

Characterizing human values is a topic deeply interwoven with the sciences, humanities, political philosophy, art, and many other human endeavors. In recent years, a number of thinkers have argued that accelerating trends in computer science, cognitive science, and related disciplines foreshadow the creation of intelligent machines which meet and ultimately surpass the cognitive abilities of human beings, thereby entangling an understanding of human values with future technological development. Contemporary research accomplishments suggest increasingly sophisticated AI systems becoming widespread and responsible for managing many aspects of the modern world, from preemptively planning users' travel schedules and logistics, to fully autonomous vehicles, to domestic robots assisting in daily living. The extrapolation of these trends has been most forcefully described in the context of a hypothetical "intelligence explosion," in which the capabilities of an intelligent software agent would rapidly increase due to the presence of feedback loops unavailable to biological organisms. The possibility of superintelligent agents, or simply the widespread deployment of sophisticated, autonomous AI systems, highlights an important theoretical problem: the need to separate the cognitive and rational capacities of an agent from the fundamental goal structure, or value system, which constrains and guides the agent's actions. The "value alignment problem" is to specify a goal structure for autonomous agents compatible with human values. In this brief article, we suggest that ideas from affective neuroscience and related disciplines aimed at characterizing neurological and behavioral universals in the mammalian kingdom provide important conceptual foundations relevant to describing human values. We argue that the notion of "mammalian value systems" points to a potential avenue for fundamental research in AI safety and AI ethics.

Povzetek: Prispevek obravnava sistem vrednot sesalcev, ki so osnova za človeški sistem vrednot, pomemben za umetno inteligenco.

1 Introduction

Artificial intelligence, a term coined in the 1950's at the now famous Dartmouth Conference, has come to have a widespread impact on the modern world [1, 2]. If we broaden the phrase to include all software, and in particular, software responsible for the control and operation of physical machinery, planning and operations management, or other tasks requiring sophisticated information processing, then it goes without saying that artificial intelligence has become a critical part of the infrastructure supporting modern human society. Indeed, prominent venture capitalist Mark Andresseen famously wrote that "software is eating the world," in reference to the ubiquitous deployment of software systems across all industries and organizations, and the corresponding growth of the financial investment into software companies [3].

Nonetheless, there is a fundamental gap between the abilities of the most sophisticated software-based control systems today and the capacities of a human child or even many animals. Our AI systems have yet to display the capacity for learning, creativity, independent thought and discovery that define human intelligence. It is a near-consensus position, however, that at some point in the future, we will be able to create software-based agents whose cognitive capacities rival those of human beings. While there is substantial variability in researchers' forecasts about the time-horizons of the critical breakthroughs and the consequences of achieving human-level artificial intelligence, there it is little disagreement that it is an attainable milestone [4, 5].[1]

Some have argued that the creation of human-level artificial intelligence would be followed by an "intelligence explosion," whereby the intelligence of the software-based

[1] There have been a number of prominent thinkers who have expressed strongly conservative viewpoints about AI timelines. See, for example, commentaries by David Deutsch, Rodney Brooks, and Douglas Hofstadter [6–8].

Sarma, Gopal P., and Nick J. Hay. "Mammalian Value Systems." *Informatica* 41.4 (2017).

Neuropsychology and AI Safety

- View human values from the perspective of neuropsychological foundations
- Suggested decomposition of human values: *1) mammalian values 2) human cognition 3) several millennia of human social and cultural evolution*

Mammalian Value Systems

Gopal P. Sarma
School of Medicine, Emory University, Atlanta, GA USA
E-mail: gopal.sarma@emory.edu

Nick J. Hay
Vicarious FPC, San Francisco, CA USA
E-mail: nnickhay@gmail.com

Keywords: friendly AI, value alignment, human values, biologically inspired AI, human-mimetic AI

Received: June 24, 2013

Characterizing human values is a topic deeply interwoven with the sciences, humanities, political philosophy, art, and many other human endeavors. In recent years, a number of thinkers have argued that accelerating trends in computer science, cognitive science, and related disciplines foreshadow the creation of intelligent machines which meet and ultimately surpass the cognitive abilities of human beings, thereby entangling an understanding of human values with future technological development. Contemporary research accomplishments suggest increasingly sophisticated AI systems becoming widespread and responsible for managing many aspects of the modern world, from preemptively planning users' travel schedules and logistics, to fully autonomous vehicles, to domestic robots assisting in daily living. The extrapolation of these trends has been most forcefully described in the context of a hypothetical "intelligence explosion," in which the capabilities of an intelligent software agent would rapidly increase due to the presence of feedback loops unavailable to biological organisms. The possibility of superintelligent agents, or simply the widespread deployment of sophisticated, autonomous AI systems, highlights an important theoretical problem: the need to separate the cognitive and rational capacities of an agent from the fundamental goal structure, or value system, which constrains and guides the agent's actions. The "value alignment problem" is to specify a goal structure for autonomous agents compatible with human values. In this brief article, we suggest that ideas from affective neuroscience and related disciplines aimed at characterizing neurological and behavioral universals in the mammalian kingdom provide important conceptual foundations relevant to describing human values. We argue that the notion of "mammalian value systems" points to a potential avenue for fundamental research in AI safety and AI ethics.

Povzetek: Prispevek obravnava sistem vrednot sesalcev, ki so osnova za človeški sistem vrednot, pomemben za umetno inteligenco.

1 Introduction

Artificial intelligence, a term coined in the 1950's at the now famous Dartmouth Conference, has come to have a widespread impact on the modern world [1, 2]. If we broaden the phrase to include all software, and in particular, software responsible for the control and operation of physical machinery, planning and operations management, or other tasks requiring sophisticated information processing, then it goes without saying that artificial intelligence has become a critical part of the infrastructure supporting modern human society. Indeed, prominent venture capitalist Mark Andresseen famously wrote that "software is eating the world," in reference to the ubiquitous deployment of software systems across all industries and organizations, and the corresponding growth of the financial investment into software companies [3].

Nonetheless, there is a fundamental gap between the abilities of the most sophisticated software-based control systems today and the capacities of a human child or even many animals. Our AI systems have yet to display the capacity for learning, creativity, independent thought and discovery that define human intelligence. It is a near-consensus position, however, that at some point in the future, we will be able to create software-based agents whose cognitive capacities rival those of human beings. While there is substantial variability in researchers' forecasts about the time-horizons of the critical breakthroughs and the consequences of achieving human-level artificial intelligence, there it is little disagreement that it is an attainable milestone [4, 5].[1]

Some have argued that the creation of human-level artificial intelligence would be followed by an "intelligence explosion," whereby the intelligence of the software-based

[1] There have been a number of prominent thinkers who have expressed strongly conservative viewpoints about AI timelines. See, for example, commentaries by David Deutsch, Rodney Brooks, and Douglas Hofstadter [6–8].

Sarma, Gopal P., and Nick J. Hay. "Mammalian Value Systems." *Informatica* 41.4 (2017).

Neuropsychology and AI Safety

- View human values from the perspective of neuropsychological foundations
- Suggested decomposition of human values: *1) mammalian values 2) human cognition 3) several millennia of human social and cultural evolution*
- Relevant disciplines include affective neuroscience, animal behavior, biological anthropology, comparative neuroanatomy, etc.

Sarma, Gopal P., and Nick J. Hay. "Mammalian Value Systems." *Informatica* 41.4 (2017).

Mammalian Value Systems

Gopal P. Sarma
School of Medicine, Emory University, Atlanta, GA USA
E-mail: gopal.sarma@emory.edu

Nick J. Hay
Vicarious FPC, San Francisco, CA USA
E-mail: nnickhay@gmail.com

Keywords: friendly AI, value alignment, human values, biologically inspired AI, human-mimetic AI

Received: June 24, 2013

Characterizing human values is a topic deeply interwoven with the sciences, humanities, political philosophy, art, and many other human endeavors. In recent years, a number of thinkers have argued that accelerating trends in computer science, cognitive science, and related disciplines foreshadow the creation of intelligent machines which meet and ultimately surpass the cognitive abilities of human beings, thereby entangling an understanding of human values with future technological development. Contemporary research accomplishments suggest increasingly sophisticated AI systems becoming widespread and responsible for managing many aspects of the modern world, from preemptively planning users' travel schedules and logistics, to fully autonomous vehicles, to domestic robots assisting in daily living. The extrapolation of these trends has been most forcefully described in the context of a hypothetical "intelligence explosion," in which the capabilities of an intelligent software agent would rapidly increase due to the presence of feedback loops unavailable to biological organisms. The possibility of superintelligent agents, or simply the widespread deployment of sophisticated, autonomous AI systems, highlights an important theoretical problem: the need to separate the cognitive and rational capacities of an agent from the fundamental goal structure, or value system, which constrains and guides the agent's actions. The "value alignment problem" is to specify a goal structure for autonomous agents compatible with human values. In this brief article, we suggest that ideas from affective neuroscience and related disciplines aimed at characterizing neurological and behavioral universals in the mammalian kingdom provide important conceptual foundations relevant to describing human values. We argue that the notion of "mammalian value systems" points to a potential avenue for fundamental research in AI safety and AI ethics.

Povzetek: Prispevek obravnava sistem vrednot sesalcev, ki so osnova za človeški sistem vrednot, pomemben za umetno inteligenco.

1 Introduction

Artificial intelligence, a term coined in the 1950's at the now famous Dartmouth Conference, has come to have a widespread impact on the modern world [1, 2]. If we broaden the phrase to include all software, and in particular, software responsible for the control and operation of physical machinery, planning and operations management, or other tasks requiring sophisticated information processing, then it goes without saying that artificial intelligence has become a critical part of the infrastructure supporting modern human society. Indeed, prominent venture capitalist Mark Andresseen famously wrote that "software is eating the world," in reference to the ubiquitous deployment of software systems across all industries and organizations, and the corresponding growth of the financial investment into software companies [3].

Nonetheless, there is a fundamental gap between the abilities of the most sophisticated software-based control systems today and the capacities of a human child or even many animals. Our AI systems have yet to display the capacity for learning, creativity, independent thought and discovery that define human intelligence. It is a near-consensus position, however, that at some point in the future, we will be able to create software-based agents whose cognitive capacities rival those of human beings. While there is substantial variability in researchers' forecasts about the time-horizons of the critical breakthroughs and the consequences of achieving human-level artificial intelligence, there it is little disagreement that it is an attainable milestone [4, 5].[1]

Some have argued that the creation of human-level artificial intelligence would be followed by an "intelligence explosion," whereby the intelligence of the software-based

[1] There have been a number of prominent thinkers who have expressed strongly conservative viewpoints about AI timelines. See, for example, commentaries by David Deutsch, Rodney Brooks, and Douglas Hofstadter [6–8].

Neuropsychology and AI Safety

SEEKING

RAGE

FEAR

LUST

CARE

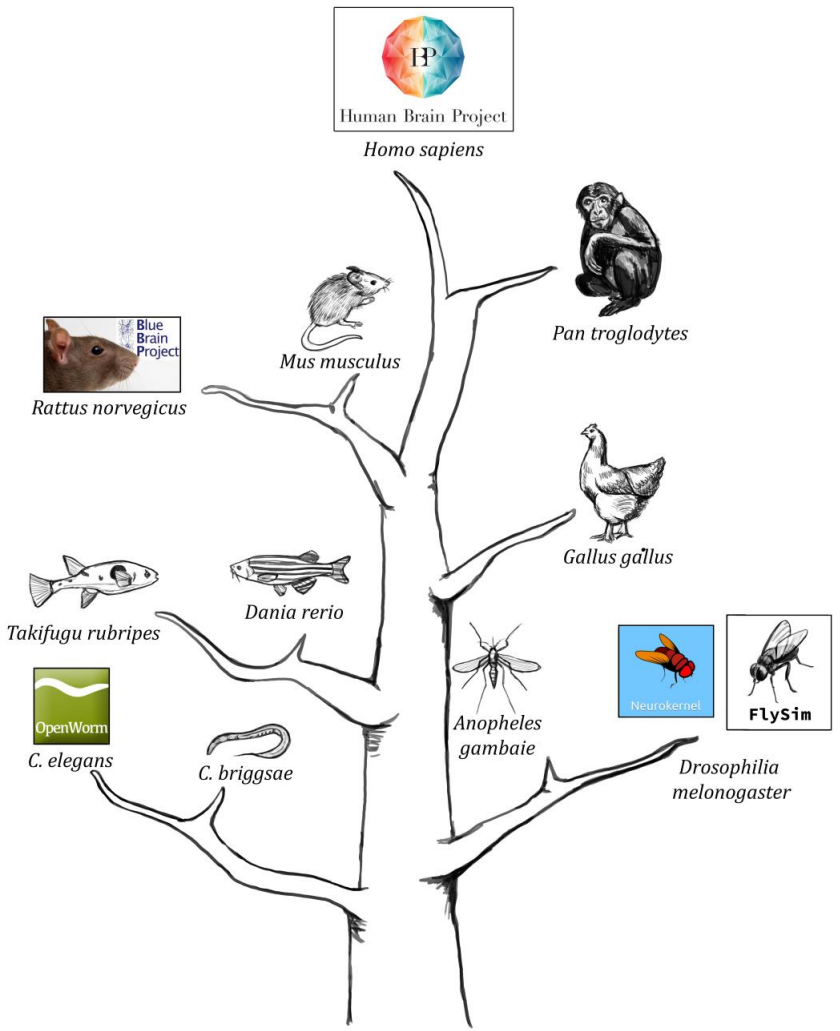
PANIC/GRIEF

PLAY

Panksepp, Jaak. *Affective Neuroscience: The Foundations of Human and Animal Emotions.* Oxford University Press, 2004.

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

Integrative Biological Simulation



NEURON

BluePyOpt

NetPyNE

Bionet

Geppetto

ChannelPedia

NeuroMLDB

Neuropsychology and AI Safety

SEEKING

RAGE

FEAR

LUST

CARE

PANIC/GRIEF

PLAY

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

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

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