Agency and Sensorimotor Systems

Response to a Call for Research Proposals, California Institute for Machine Consciousness (CIMC)

Linas Vepštas$^{[ORCID\quad 0000-0002-2557-740X]}$

BrainyBlaze Dynamics, OpenCog Foundation <linasvepstas@gmail.com>

Abstract. The core idea of consciousness as a form of embodied cognition is popular and widespread. Yet an abstract mathematical description of how this might work, anchored to first principles in physics, is absent. A sensory system "perceives" the world: but what is it that is perceived? It can't just be some "data" or "message", as that seems to require a data–definition format and some homunculus to define it. Perhaps sensation is movement of structural relationships across a barrier or boundary between the agent's finite and limited self/inside and the unbounded outside of the external world. This requires the barrier only to limit what gets across. What is the mathematical description of such a barrier? Perhaps the "self" is an accretion of structure inside an agent. How can this accretion be formally described? Jellyfish seem to be aware of "self" and "other"; discussions of consciousness often presume such awareness. Is there an algebraic formulation for such (self-)awareness? Suppose such a formulation can be created; is there anything within it that has the form of "qualia", or is it mechanical and dead inside?

Research Objectives

Computation requires the articulation of a system using algorithmic language. Algorithms require precise structural definitions. Structural definitions are generically algebraic. Foundations for algebra include set theory, category theory and topoi. Thus, a "computational perspective on consciousness" fundamentally requires (seems to require) a reductionist description of the structure of the exterior world, the barrier that separates it from the interior world, and how that interior interacts and moves through the external world. The interior itself may have hierarchical structure: the inside of a biological creature is not a uniform puree of pancake batter, but also contains substructures. It seems this recursive structuring continues "all the way down", at least as far as biochemistry,[1] and perhaps much farther.[2]

The goal of the research is to articulate what it means for there to be an inside, an outside, and a permeable boundary between the two. Does this boundary resemble a

[1] See the review article: "All intelligence is collective intelligence", Falandays *etal*. Journal of Multiscale Neuroscience Vol 2 no 1 169-191 (2023) https://researchportal.helsinki.fi/en/publications/all-intelligence-is-collective-intelligence

[2] See "Ingressing Minds: Causal Patterns Beyond Genetics and Environment in Natural, Synthetic, and Hybrid Embodiments", Michael Levin https://osf.io/preprints/psyarxiv/5g2xj_v1

domain wall in a ferromagnetic system (*a la* Ising model)? Just as petroleum does not flow through rock, until there are enough holes/fractures in the rock allowing it to flow, exhibiting a phase transition at a critical fracture–ratio, then perhaps intelligence is similar: absent, until there is a critical level of organization? Many systems in nature exhibit self–organized criticality, and recent neuroscience results provide some indication that human intelligence operates at the point of self–organized criticality, with long neurons modulating the avalanching behavior that is characteristic of self–organized systems.[3]

One can imagine that it is easy to write a mathematical description of inside, outside and a barrier between them. It is hardly clear how to take such a system, and have it drive itself into a self–organized critical point, exhibiting any of the basic phenomena associated with intelligence.

Methodology

My personal predilection is for a rather conventional laboratory experimental approach. Forward progress is achieved thorough a mixture of theoretical articulations and experimental realizations. Try to build a system that "works". Collect the data. Analyze the data. Try to understand "what happened".

The "lab" here is software. Software is commonly distinguished into imperative vs. functional programming styles. Off in a neglected corner is the descriptive or declarative style. The declarative style splits computation into two pieces: a description of "what is", and a distinct component of "how it interacts". The description of "what is" can be understood as a description of structure, of network relationships, of graphical vertexes and edges (or dots and arrows, for category theory/topoi.) The "how it interacts" portion is a dynamical system or driver that mutates structures and relationships over time. The driver is more–or–less structure agnostic; the driver rearranges the descriptive/declarative program through combinatoric forces, e.g. free energy principles, maximum entropy principles, and/or related non-equilibrium thermodynamical principles, including conventional ideas formulated *a la* Gibbs and Boltzmann, Markov/Bayesian blankets, *etc.* That is, the driver is presumed to accord well with generally–accepted conventional concepts.

Accepting this split between declarative expressions of structure and drivers of dynamical change requires a framework *aka* language for declarations. The framework to be used is Atomese[4] Note that there are two distinct versions of Atomese; one that is an evolution of "classical Atomese", and another, developed from scratch, in Ben Goertzel's Hyperion system, which is sometimes called "Atomese 2.0". For practical and technical reasons (including stability, debugability, scalability and performance), the original formulation of Atomese will be used.

[3] See "Complex harmonics reveal low–dimensional manifolds of critical brain dynamics." Deco, *etal.* https://journals.aps.org/pre/abstract/10.1103/PhysRevE.111.014410

[4] See https://wiki.opencog.org/w/Atomese

Expected Outcomes

An experimental approach generally precludes fireworks and stymies hyperbolic claims of grand achievements. There's a certain rhythmic daily grind involved in constructing, running and analyzing.

One of the simplest sensori–motor systems is that of walking through a file–system, and "observing" what is there. This is a very different domain than the conventional 3D abstraction of Minecraft, virtual worlds or robotics. It is also distinct from the 2D pixelated ARC Prize ARC-AGI-2 challenge. Nor is it one–dimensional, in the way that streams of words/text are treated in LLM's and transformers. The hierarchical structure of a file system seems closer to the structure of "knowledge in the abstract" or "information in the abstract", and thus may provide a better experimental domain for how a knowledge–processing system ingests, manipulates, excretes, reorganizes and retains structural relationships. (Atomese is itself hierarchical tree–structured.)

The expected object of study is a self–organized collection of structures that exhibit navigation and movement through its environment (for example, a file system), together with interaction with and mutation (manipulation) of the environment.

The expected outcome is the formal (algebraic, mathematical, algorithmic) description of such a system and a characterization of its behavior in statistical and observational terms. This includes scaling behaviors, measures of turbulence, observation of avalanching behavior, movement, localization, *etc.*

Timeline

This project is already ongoing, with bits and pieces scattered across four or five distinct github sites.[5] Much of the effort so far has been an articulation of "what the problem is", and attempts to formulate it in functional code. The build–test–retry cycle is short, on the scale of weeks. This is a small project; there is no plan to create a plan, with project planning time–lines or Gantt charts. The goal is to just "move forward", as best as can be done.

Budget Justification

Hardware compute resources have already been procured, and so hardware expenditures of less than $10K are expected. The nature of this research does not seem to require the use of, or interaction with LLMs, GPT, or any of the deep–learning neural net systems; thus a budget of zero for LLMs is anticipated.

[5] These include:

- https://github.com/opencog/sensory
- https://github.com/opencog/motor
- https://github.com/opencog/evidence

The content at these sites is tagged "version 0.1" and consists primarily of design documents and a smattering of code, demos, unit tests and examples.

Some of the contemplated theoretical models might run much faster if ported to GPU's. Perhaps the theory may indicate that the descriptive, declarative systems resemble TensorFlow. If so, then a contingency must be set aside for rental of cloud–compute resources. However, the need for this, and the nature of this need is entirely unclear. There's a shadow of a hint at the edges; how it comes into play remains unknown. The current work is a low–level exploration of theory, and has not reached the stage of being compute–hungry.

The budget is primarily that of the salary for the lead researcher.

Team and collaborators

Team of one: myself: Linas Vepstas. The research above is sufficiently arcane, and the description of it is sufficiently obtuse that I have not been able to attract interested parties or collaborators. I seem to be operating in a corner of the noosphere that gets no visitors. So it goes.