POC Applied: Physics of Mind

The Goal:

Epsilon-machines and their associated algorithms are interesting data structures & algorithms in of themselves. The purpose and motivation of my project is to investigate what use can come of imposing additional structure on epsilon-machines. Specifically, what were to happen if I took observations of highly nonlinear dynamical systems, derived the corresponding epsilon-machine model of a process on it, and then imposed additional heuristics on the causal states of the epsilon machine based on states' interpretative semantics? What useful things could this extended model say?

More concretely, I want to look at machines that learn in a "self-aware" manner. How would a machine "think" about a dynamical system in relation to how it calculates epistemological causal states from physical causal states, and, in particular, how could it ascribe relationships between epistemological causal states and physical ones? Answering these questions could be the first step in a long walk of understanding at a high-level how a learning machine of a certain type systematically tends to look at and interpret information. What biases will it have, and how will its learning be affected? What types of semantic understanding can these "self-aware" learning machine accrue?

To make these questions well-posed questions, it is necessary to supplement the epsilon-machine [the casual states & the transitions] with semantic and logical information. Guided by [figuratively] intuitive principles of what it means to be "self-aware," I intend to construct a class of these supplemented machines—what I call pondering machines, or p-machines—from the more general notion of epsilon machines and topics from various areas of mathematics, physics, and philosophy [proof theory, constructive reverse math, polymodal logic, information theory, category theory, POCI, naturalism]. All this to ask the topical question encompassing all the aforementioned ones: how does a pondering machine evolve its knowledge? –Just kidding there is one more question; the last and most important thing I want to ask is how does this framework I have created change the way we think philosophically about conscious thinking in general (and what does it say about my favourite philosophy subject "Digital Ontology")?

The Systems:

There are two [ideally: artificially separated] dynamical systems of note in this maths experiment. The first one is the autonomic physical phenomena model all around the thinking machine—its physical world, so to speak, or an omnipresent *where* that it can observe things from. This dynamical system's dynamic is some physics, which in our [we human beings'] case often has nonlinear behaviour; the system is a particular universe where things happen over time, and the states are the physical states of the system at given time.

The more interesting one is the second dynamical system. This is the one's model is "executed" by the "brain" of the thinking machine. How does it associate in its mind a probability to the chance of something happening or "being real," and what errors are incurred compared to the actual probability of that? The system in this case is the body of knowledge of the thinker; the dynamic is one that takes observations of a physical system to an [extremely nonlinear] abstract idea space with associated epistemological probability, logical, and semantic structure.

The Mathematical Mechanism:

First I show that, if we turn on logical dependence between causal states, Prof. Crutchfield's predictive equivalence relation is a refinement of the predictive equivalence relation used in Gentzen-style proofs. Specifically, the epsilon equivalence relation lets us talk about modal truths. Instead of something being 'True' or 'False', we can allow ' X% Certain' values, and these are taken into account by the probability distributions in the epsilon machines themselves. Interpreting logical dependences between causal states also endows our epsilon-machine diagrams with new graph-theoretic depths, corresponding to the proof-theoretic "strength" of a state. From here, I then create a proof calculus in the style of Giorgi Japaridze's "cirquent calculus."

This new mixed-state-diagram-esque graph now has two modal equalities to speak of. First, it is predictive in terms of epistemological probabilities, and second it's predictive in terms logical dependence (equivalently, in terms of directed graph topology). Additionally, in order to make calculable comparisons, I introduce a simple semantic scheme I call "Platonic dictionaries," which allows one to match a well formed logical thought to an equivalence class of logical constructions. Since these equality modalities are independent and hierarchical, we can consider one as an intensional equality and the other as an extensional equality; i.e we can define for any "history" morph 4 [but actually only 3 useful] forms of "predictive equivalence"—'Not Equivalent', 'Equivalent', 'Identical', and 'Canonical'. There is more to the pondering machine construction (and much more to be said about its implications), but with these bare essentials, we can start thinking of our state transitions in terms of both proof theoretic strength as well as probability distribution. (Awesomely enough, we can also turn on abstract resource management if we consider nominal equivalence relations, but that is outside the scope of this project.) The edges are labelled much like we would expect

with an "X | Y", where the only difference is: here one of X or Y is (explicitly) an *epistemological* probability and the other is, not a symbol, but a construction [i.e an algorithm or Constructive axiom] of the witness [in the reverse mathematics/proof-theoretic sense] which leads from one causal state to the next.

What in the world does one of these pondering machine diagram representations actually look like? I have pictures! The most notable topological lemmas from the construction of these p-machines we'll be discussing at length are: 1) are every "subgraph" of the p-machine diagram representation is itself an epsilon machine; 2) By application of Gödel's second incompleteness theorem, the roots of the graph are by necessity at least partially non-constructive [there's a very interesting discussion to be had on that here]; 3) There is exactly one stationary region in all p-machine proof graphs [the diagrams], and specifically the one that it is [which is constructed in the proof] has fascinating philosophical implications about thinking in general. Bonus lemma: 4) All learning processes, *in a very [*figuratively*] intuitive sense* (this one's pretty hand-wavy/non-rigorous), can be reduced to the exact algorithm detailing the systematic construction of a p-machine, and thus the p-machine construction [as of now called the "p-learning process"] is robust enough to model all time-evolved self-aware learning mechanisms.

Appropriateness:

I ensure that this project has interesting things to by doing three things. 1) explicitly constructing something new, 2) avoiding collection/set-theoretic questions (as they would need us to say something about the internals of the casual states) and shifting instead to a category-theoretic perspective which requires we only need be sure about the relationship between casual states (this prevents us from confounding what is *causal* in the learning process and what is *determined* by the learning process), 3) by proving the important, provoking bits Constructively (this last one entails we'll always have a concrete object we can talk about when thinking about examples). Ultimately, this paper is a philosophical application of the mathematics I learned in my background and in POCI, and while there is interesting math involved, I am not setting out rigorously prove the whole framework (yet—not for this project anyway). The paper and the p-machine as described here will be more intended to start for my fellow classmates thinking and discussions based on the cumulative insights that will have been made in this project.

Hypothesis:

Having submitted my proposal a week late, i.e, having already obtained some significant preliminary results, I am certain I'll have at the very least a couple of interesting things to say about: some behavioural patterns of and the limits of "conscious" learning, pancomputionalism, naturalism in mathematics, and the potential future of "Strong AI's" & the process how they may one day be given personalities. These will be controversial on a philosophical level, but I intend to convince you all with the mathematics and intuitive reasoning!

Time & Steps:

0. Write this project proposal – 0 days (I just wrote it)
1. Write a small portion of my report just concerning the explanation and construction of the pondering machines – 8 days
2. Explore obvious preliminary lemmas and facts about the construction, prodding it with different maths, such as graph theory, category theory, proof theory, and information theory [especially the work of Milne.] – 5 days
3. Construct a finite step-through of an actual learning process a pondering machine might have actualized to learn about the different foundational models of arithmetic (up until the construction of the surreal numbers) [a concrete example of my learning process applied]; mark observations. – 13 days
4. Write up the class presentation slides – 1 day
5. Write the final report – 5 days

Total days: 32 (and I already done about 4-5 days' work)

Thus, **yes, I can complete the project in time, if approved.**