Goal:

I propose to study a simple machine model of an organism in its environment, with an emphasis on demonstrating some kind of "cognitive" behavior of the organism beyond mere response to stimulus. The agent and environment can be described as two coupled nondeterministic transducers, with the agent machine reading the symbols generated by the environment machine and vice versa, together forming one overall system whose state comprises the state of both machines.

Starting from this basic framework for the agent/environment interaction, several interesting directions are possible. My interest is in investigating the extent to which the agent "controls" and "models" its environment. My conceptualization of what this will mean is still rather open-ended, but I intend to measure quantities such as the mutual information between the agent's present and its past or future, as well as the "surprise" generated by the environment from the agent's point of view.

The meat of the project may ultimately lie in investigating several different cases -- different types of agents inhabiting different types of environment -- and the consequences of qualitatively different agent or environment architectures.

System:

The system will consist of a simple model of an agent or organism interacting with its environment, with agent and environment each "transducing" the symbols of the other. The environment machine, and possibly the agent machine as well, will be nondeterministic. The state space comprises two general state variables: the internal state of the agent, and that of the environment. The dynamic arises from the state transitions of each machine relative to the symbol it has just read from the other.

Dynamical properties:

I expect to find that "metastable recurrent sets" of the agent machine will be of great importance -- that is, state regimes that function as the recurrent set of the machine relative to most input strings, save a few that can "disrupt" the agent's behavior and push it into another metastable regime. In the simplest case, such a set can consist of only one state -- and indeed, any set that is recurrent under some inputs satisfies this property in a trivial sense. We are interested, however, in those sets for which recurrence is a typical norm rather than an incidental possibility. The environment machine may be designed to show similar wells of stability, which are only vacated via some low-probability event, which may be of interest as well.

These dynamics may merit a deep investigation and I may need to draw very widely from the theory of dynamical systems and statistical mechanics to describe them adequately. This is the sort of thing that could be a backup project topic in itself if my current one turns out to be too broad or too ambitious.

Intrinsic computation properties:

My emphasis is on the computational properties of the agent's internal state, which can be construed as its "percept" that structures its behavior in its environment. The computations that give rise to it and the computations that result from it will be the focus of the project.

Methods:

I'll build something in SAGE that can implement the systems that I'm interested in. Rather than constructing the machine representation of the system by hand, it may behoove me to build an "emergent machine" -- a system encompassing various quantities that change state in discrete time, with any continuous quantities being coarse-grained at some stage in representation.

Steps and Time:

-design agent and environment machines - 10-15 days

-write code implementing the agent/environment system - will be done in the same block of time as the above

-read literature - 7-14 days, also overlapping with the above.

-calculate quantities, draw conclusions - up to 5 days

-write report - up to 5 days

I believe I can complete the project within one month.