Internal State Predictability as an Evolutionary Precursor of ***Self-Awareness*** and Agency

Jaerock Kwon, Yoonsuck Choe

Neural Intelligence Lab

Computer Science Department

Agenda

- Research Motivations
- Introduction
  - Self-awareness
  - Internal State
- Methods
- Experiments and Results
  - Neuroevolution
  - Time series prediction
- Conclusion and Discussion

Research Motivations

- Why are we conscious?
  - What brain activities?
  - What kind of evolutionary pressure?
- It is too intricate to answer
- An alternative way to investigate
  - Necessary conditions for the emergence of self-awareness, a primitive form of consciousness

Self-awareness

- Self-awareness
  - has an important role in cognitive processes [Block 1995]
- Task performance
  - An agent doe not necessarily have to be self-aware

- Then, why have intelligent agents evolved to have self-awareness?

Approach

- The attributes of self-awareness
  - Still uncertain [Taylor 2007]
- So, the emergence of self-awareness
  - It is difficult to track down
- One way to circumvent the problem
  - Find necessary conditions for the emergence
  - Assess their evolutionary value

Internal State and Sense of Self

- Modeling of sensory motor dynamics
  - The central nervous system (CNS)
    - models sensory motor dynamics
    - The model seems to reside in the cerebellum [Wolpert, Miall, & Kawato 1998]
- Exploring one's internal state
  - can lead to a sense of self
  - The sense of self
    - maybe a prerequisite to build a machine with consciousness [Kawamura *et al*. 2005]

Internal State

- Neuronal activation levels
  - can be considered as the *state* of a neural system
- The state of a neural network
  - the current activation levels of the hidden units [Bakker & de Jong 2000]
- The system state
  - could be viewed as consciousness in a way [Rolls 2007]
- Physiological arguments
  - The firing rate of each neuron in the inferior temporal visual cortex tells stimuli applied to the cortex [Rolls 2007]
  - Spiking activities from place cells in the hippocampus can be used to rebuild certain features of the spatial environment [Itskov & Curto 2007]

Internal State Predictability (ISP)

- The predictability of one's own internal state trajectory.

- Our results show
  - ISP has a strong impact on performance of the agents
  - ISP could have led intelligent agents to develop self-awareness

In summary,

- Spiking patterns of neurons
  - One’s internal state
- Knowing internal state of oneself
  - The first step of being conscious

Agenda

- Research Motivations
- Introduction
  - Self-awareness
  - Internal State
- Methods
  - Neuro-evolution
  - Time series prediction
- Experiments and Results
- Conclusion and Discussion

Method

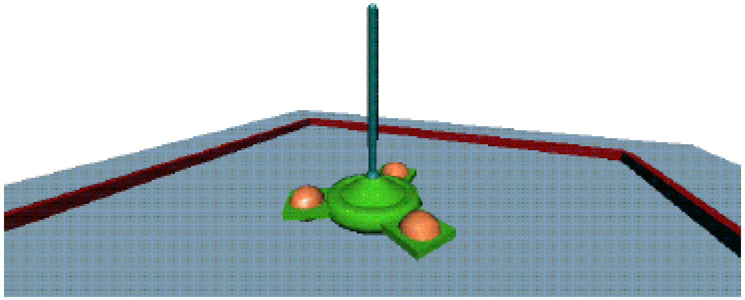
- Understanding one’s own internal state (self-aware or consciousness)
  - Knowing what is going to happen in one’s own internal state
- Quantified such an understanding
  - as the predictability of the internal state trajectory
- Evolutionary value of such an understanding?
  - We evolved sensory motor agents
    - with recurrent neural network controllers

Method

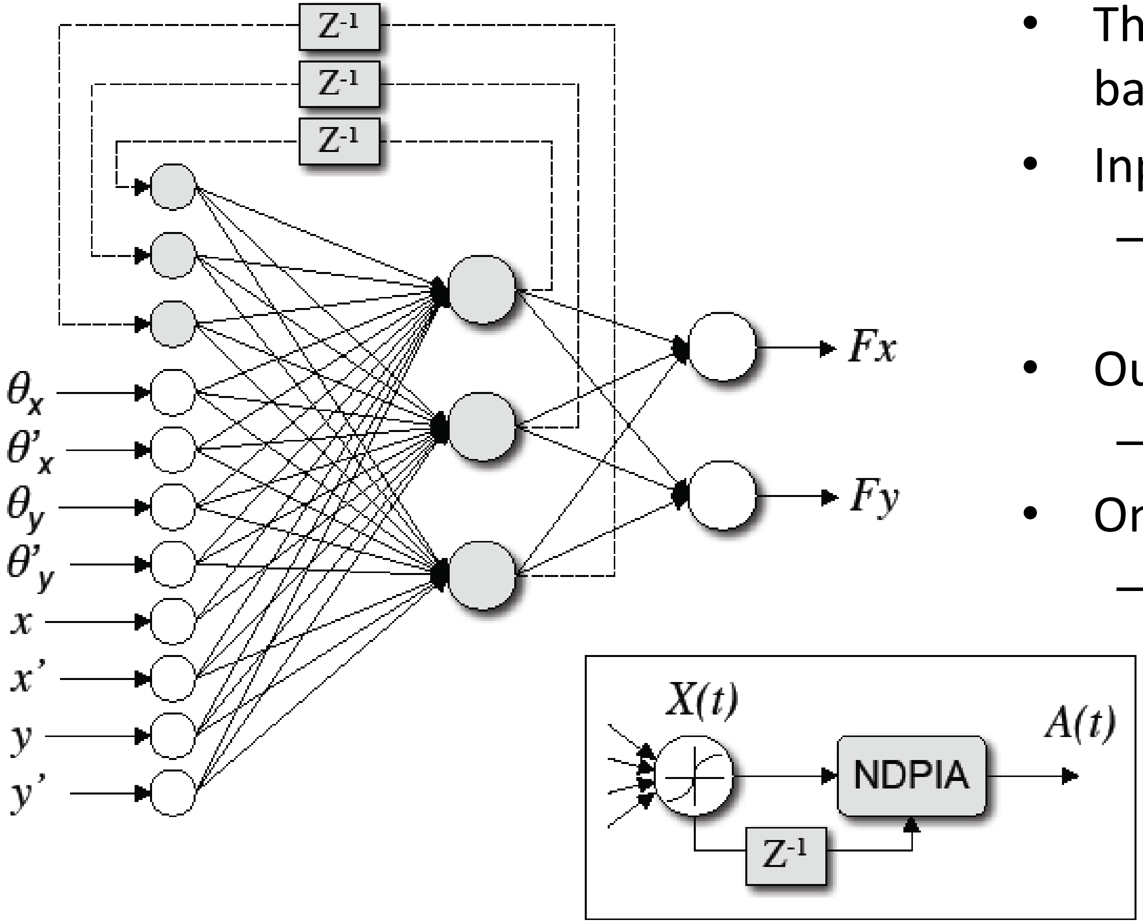
- Task
  - 2DOF pole balancing
- Training the controllers
  - Neuro-evolution
- The neural activity in the hidden layer
  - The internal state of an agent
- The predictability of the neural activity
  - Measured by a supervised learning predictor

2DOF Pole Balancing

- A cart with a pole moves in a plane
  - Balance the pole as long as possible
- Why 2D pole balancing?
  - Easy to understand and visualize
  - Embody many essential aspects of a whole class of learning task



Recurrent Artificial Neural Network



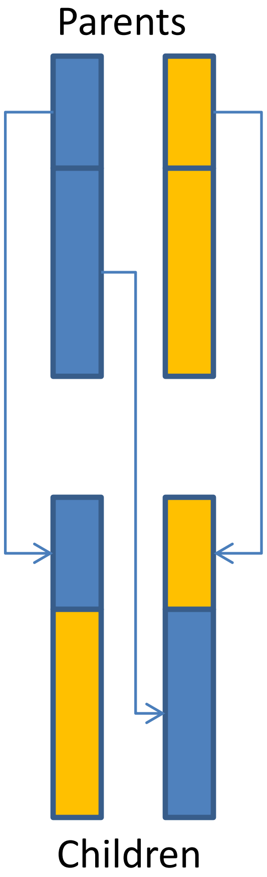
- The controller of a pole balancing agent
- Inputs neurons (8)
  - Pole velocity and acceleration of x and y positions and angles
- Outputs neurons (2)
  - Force toward x and y directions
- One hidden layer, three neurons
  - Recurrent to the input nodes

Genetic Algorithm

- Evolution
  - The changes seen in the inherited traits of a population from one generation to the next [Wikipedia]
- Genes
  - Pass to offspring during reproduction
- Reproduction
  - Recombination of genes
  - Not perfect
- Natural selection
  - Inherited traits that are helpful for survival and reproduction become more common in a population

Genetic Algorithm

- A population of *abstract representation* of candidate solutions
  - The abstract representation: chromosomes (genomes)
  - Evolve to have better solutions
  - The evolution starts from a population of randomly generated individuals
- Natural selection
  - In each generation, every individual is evaluated based on *fitness*
- Reproduction
  - Generate a second generation population
  - Recombination: crossover
  - Mutation



Neuro-Evolution

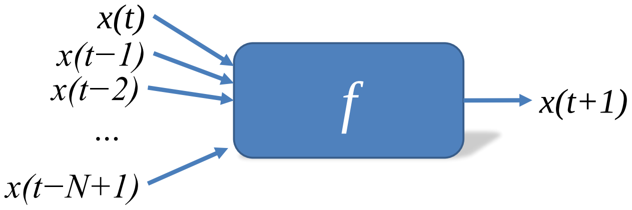
- Nonlinear control system
- The neural networks were trained by GA
- Network connection weights were evolved to balance the pole
- Chromosome / genome
  - A series of all the network weights
- Fitness function
  - The number of pole balancing steps

Agenda

- Research Motivations
- Introduction
  - Self-awareness
  - Internal State
- Methods
  - Neuro-evolution
  - Time series prediction
- Experiments and Results
- Conclusion and Discussion

Time Series Prediction

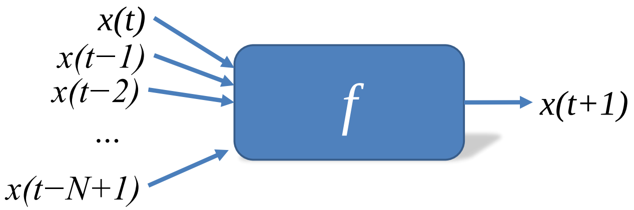
- Time series
  - A sequence of data from a dynamic system
- The activation level of hidden neurons
  - Can be considered as a time series
- Time series prediction

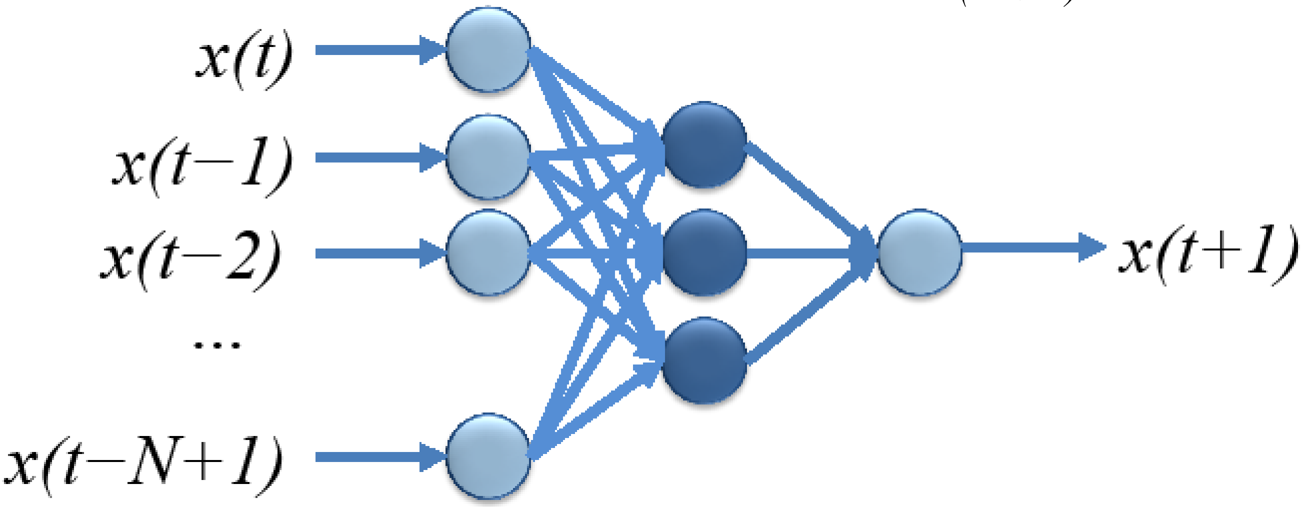


$$x(t+1) = f(x(t), x(t-1), x(t-2), \ldots, x(t-N+1))$$

Neural Network Predictor

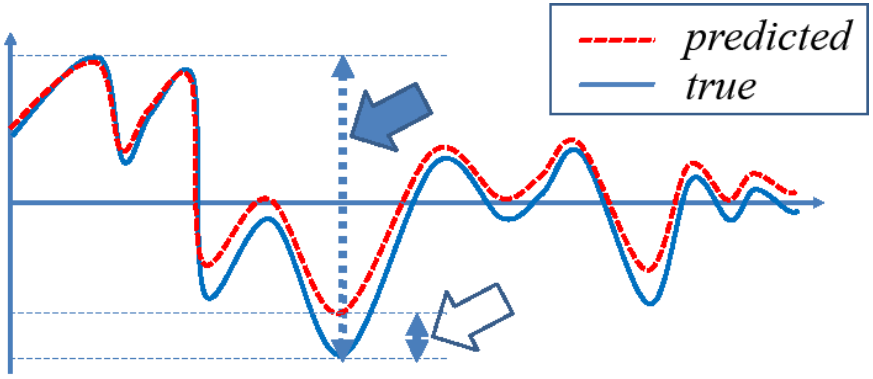
- Feed forward neural networks have been widely used

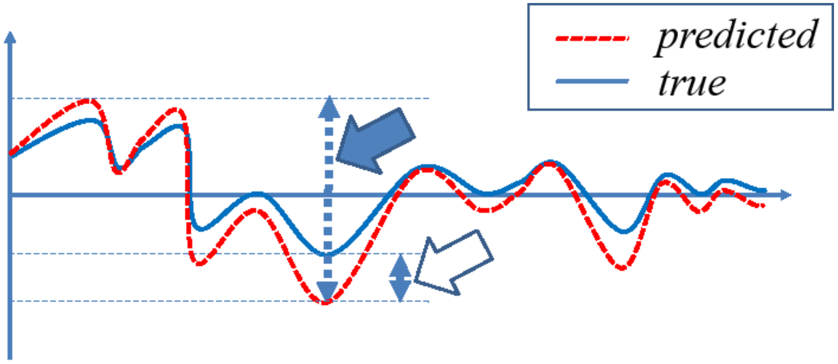




Adaptive Error Rates

- Error in forecast a future state
  - Should be adapted to the envelope of activation





Agenda

- Research Motivations
- Introduction
  - Self-awareness
  - Internal State
- Methods
  - Neuro-evolution
  - Time series prediction
- Experiments and Results
- Conclusion and Discussion

Training the Controllers

- Pole balancing agents with a recurrent neural network
- The networks were trained by genetic algorithms
- Force to the pole between -10N and 10N was applied at 10 millisecond intervals
- The pole length : 0.5 meter
- The initial condition: 0.01 radian tilted from x-z and y-z plane respectively
- The area where the cart moved was 3 x 3 m^2

Neuro-Evolution

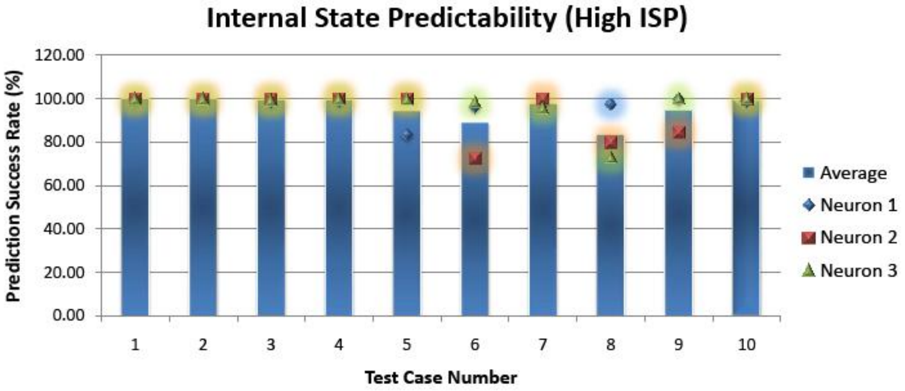
- Fitness
  - The number of steps where a network was able to keep the pole within ±15 degree
- Parameters
  - Population size: 50
  - Mutation rate: 0.2
  - Crossover rate: 0.7
  - Desired steps of pole balancing: 5,000
- Get around 130 successful networks

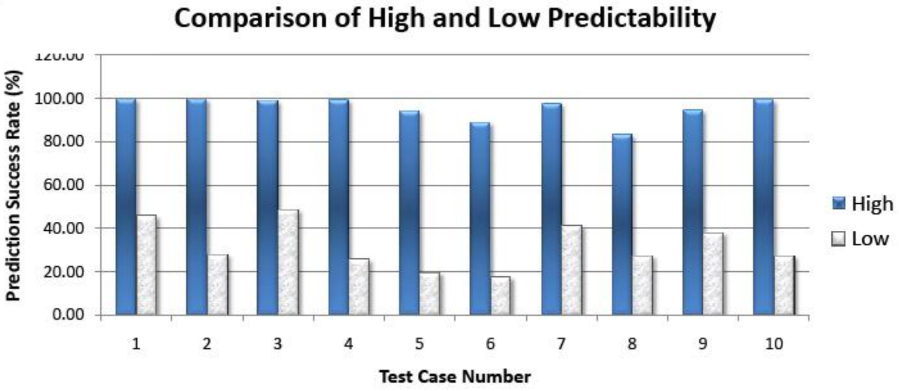
Training the Neural Network Predictor

- ISP can be measured using a feed forward neural network predictor
- The predictor quantifies the predictability of three hidden neurons' outputs
- The size of sliding window: 4
- Using 3,000 activation values
  - Training / test : 2,000 / 1,000
- Back-propagation algorithm
  - Learning rate : 0.2

Performance Measurement

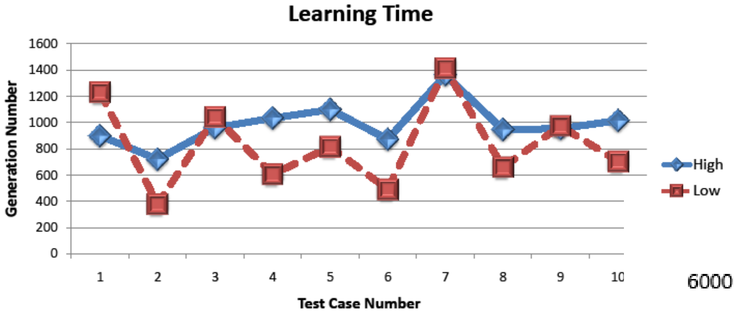
- Choose top-10 ISP networks and bottom-10 ISP
- Most of top-10 ISP networks show 99% of prediction rate
- Most of bottom-10 ISP networks show 17.37% to 48.53%

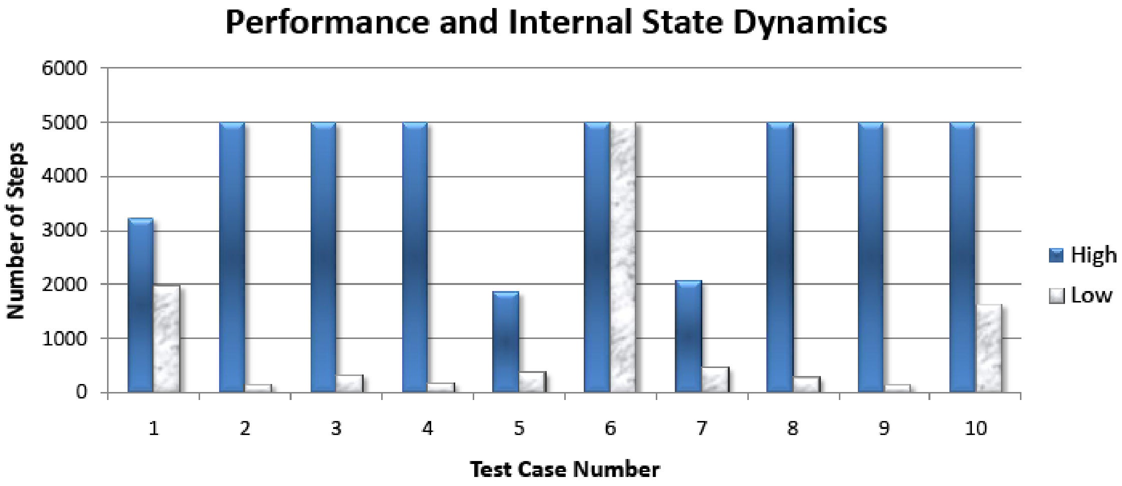




Performance Measurement

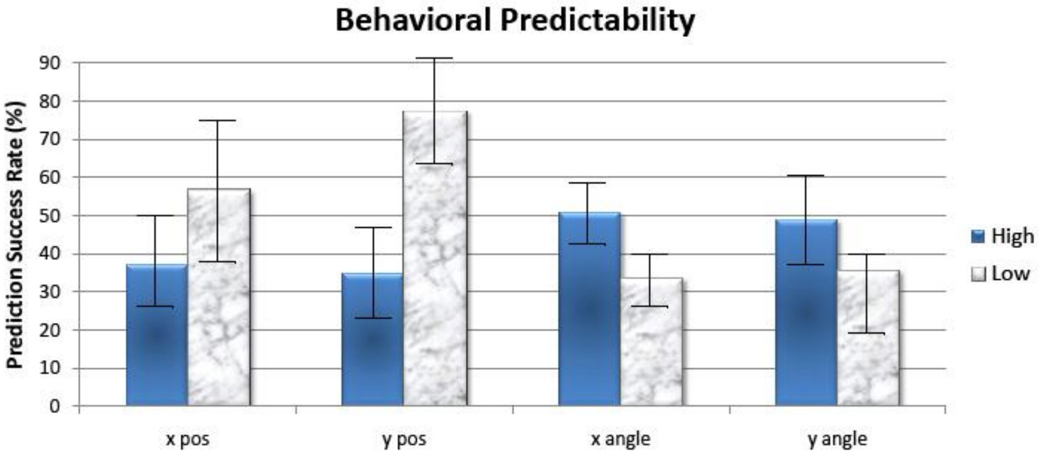
- Compare performance of two ISP groups
- Make the initial condition harsher
  - 0.07 radian to x-z plane, 0.04 radian to y-z plan



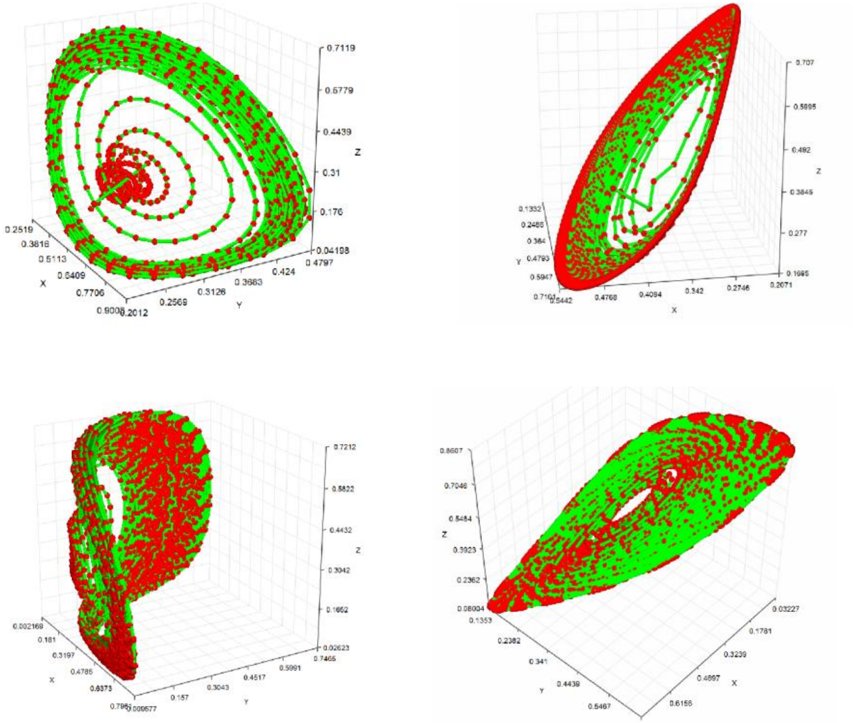


Behavioral Predictability

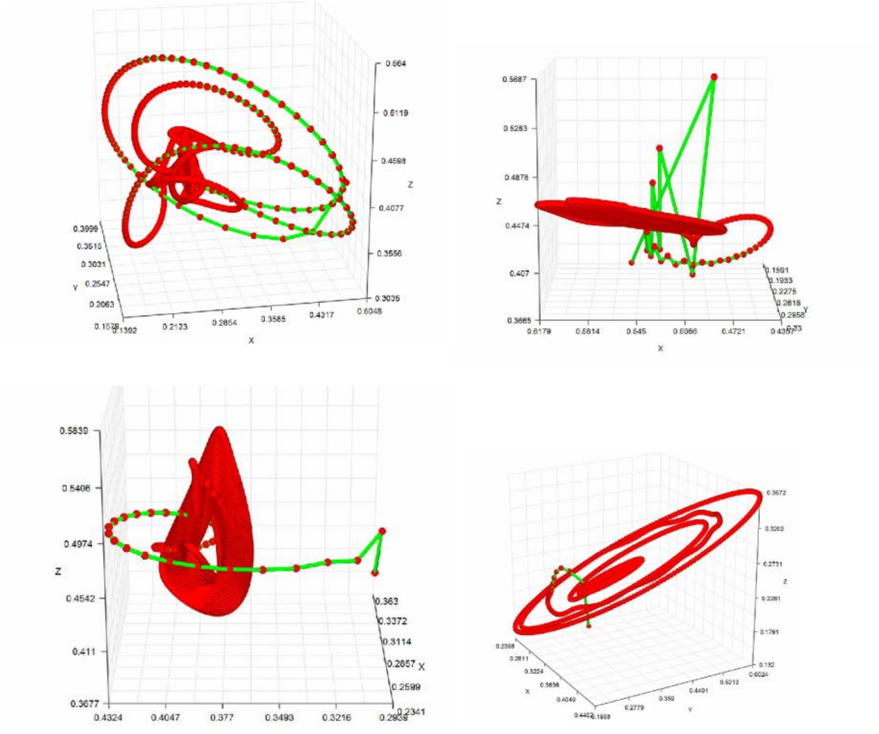
- Do simple internal state trajectories reflect behavioral properties?
  - – Seems no



Examples of Internal Dynamics

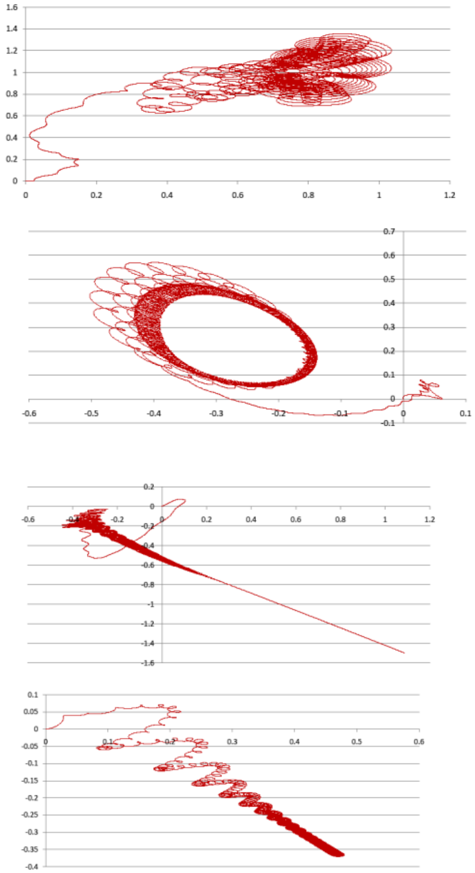


High ISP group

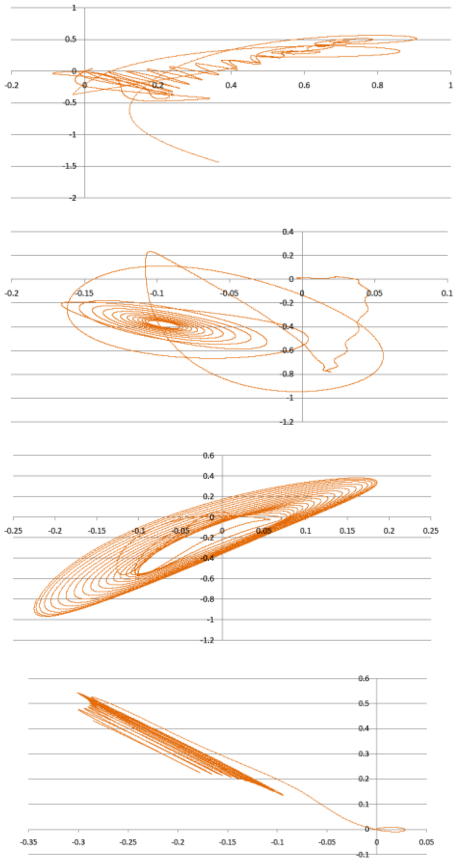


Low ISP group

Example of Behavioral Trajectories



Position trajectories in the high ISP group



Position trajectories in the low ISP group

Conclusion and Discussion

- Starting with individuals showing same behavioral performance
- More predictable internal dynamics
  - achieved higher level of performance in harsher environmental conditions
  - may have a survival value in evolutionary context
- Internal properties can affect external behavioral performance in changing environments
- The results show how **an Initial stepping stone** to **self-awareness** has been formed in the evolutionary pathway