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

ICDL 2008

August 11, 2008

Jaerock Kwon and **Yoonsuck Choe**

Department of Computer Science

Texas A&M University

Motivation

The concept of self (self-awareness, agency) is an important yet hard subject:

- It may lead to consciousness.
- It may be necessary for social interaction.
- It may play an important role in cognition (Block 1995).

Research Question: Self-Awareness

Why did self-awareness (or the sense of self) evolve?

- Self-awareness is an internal state that may be transparent to the process of evolution (cf. high-performance zombie).
- This is a hard question to answer without getting tangled in philosophical debate.

Strategy: Investigate the **necessary condition** of self-awareness that may be less controversial.

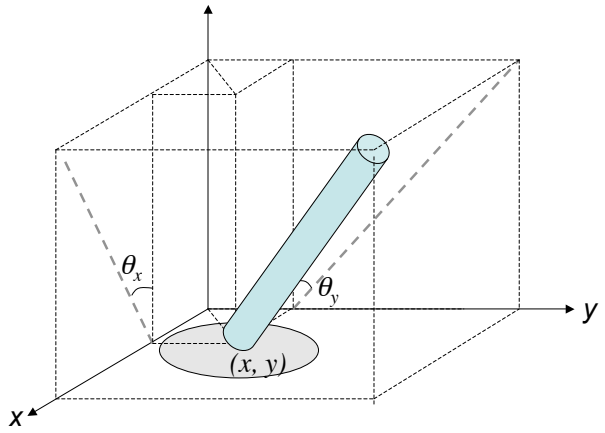
Approach

Identify **necessary conditions** of self-awareness:

- Sense of self and agency are closely related.
- **Authorship** is a key ingredient: *"I" prescribe my actions, and "I" own them*.
- Important property of authorship: My actions are **highly predictable** while others' are not.

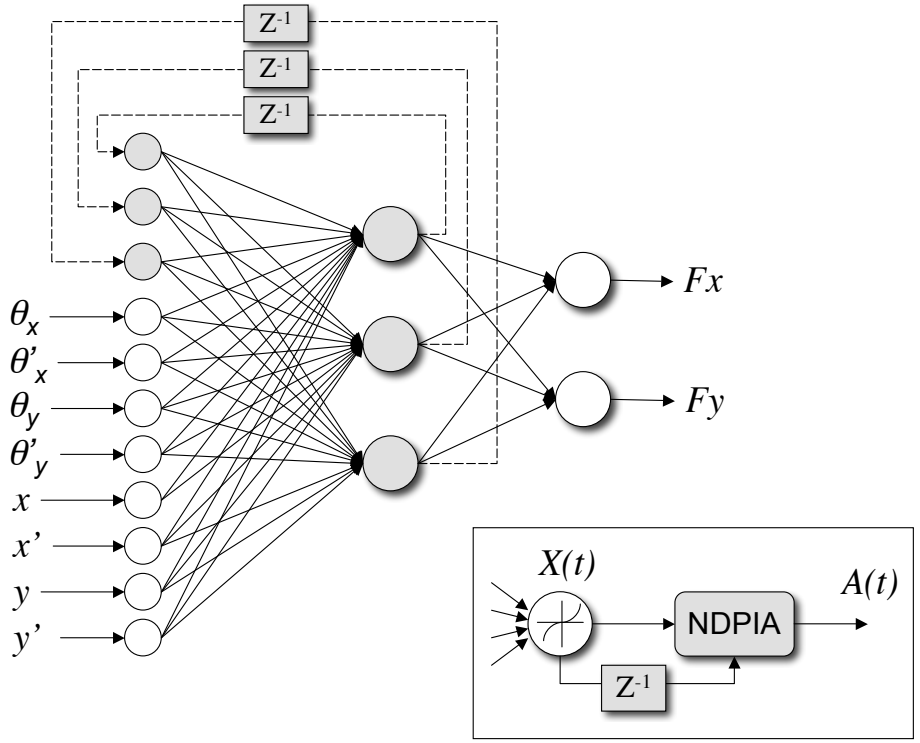
Necessary condition identified: Need to be able to **predict** one's own internal state (cf. Nolfi et al. 1994).

Method (Task): 2D Pole-Balancing



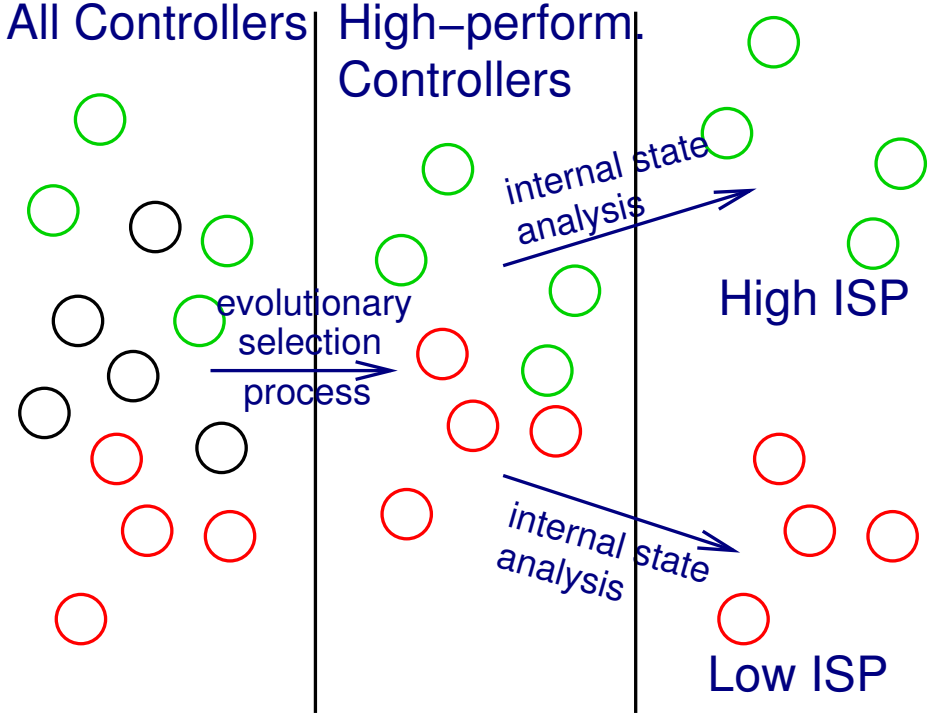
- Physical parameters of the pole balancing system: position (x, y); velocity $(\dot{x}, \dot{y})$; pole angle $(\theta_x, \theta y)$; angular velocity $(\dot{\theta_x}, \dot{\theta_y})$.

Method: Neuroevolution Controller



- Recurrent neural network for 2D pole balancing.
- Trained with standard neuroevolution.
- Investigate the internal state trajectories.

Sketch of the Method

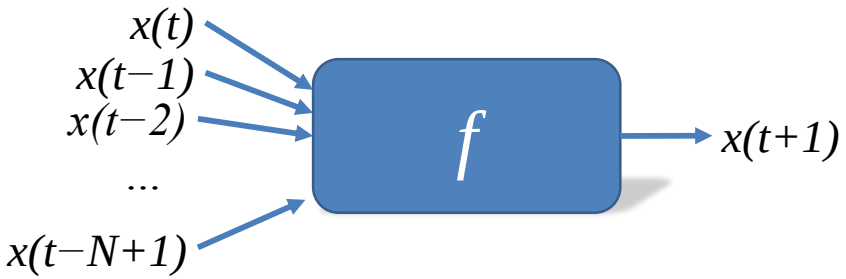


1. Evolve controllers to meet a fixed performance criterion (fitness does not measure predictability) in pole-balancing tasks.
2. Group high-performance individuals in to high- and low internal state predictability (ISP) groups.
3. Test the two groups in harder tasks.

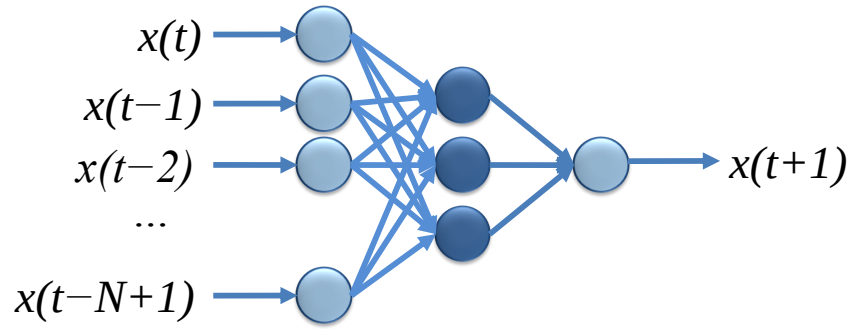
Method: Experimental Setup

- Neuroevolution:
  - population size 50
  - mutation rate 0.2; cross over rate 0.7.
- 2D pole balancing task:
  - Pole should be balanced within 15° within a 3 m $\times$ 3 m arena.
  - Force applied to cart every 0.1 second (= one step).
  - Success if pole balanced over 5,000 steps.

Method: Measuring Predictability



$$\hat{x}(t+1) = f\left(x(t), x(t-1), x(t-2), \cdots, x(t-N+1)\right).$$

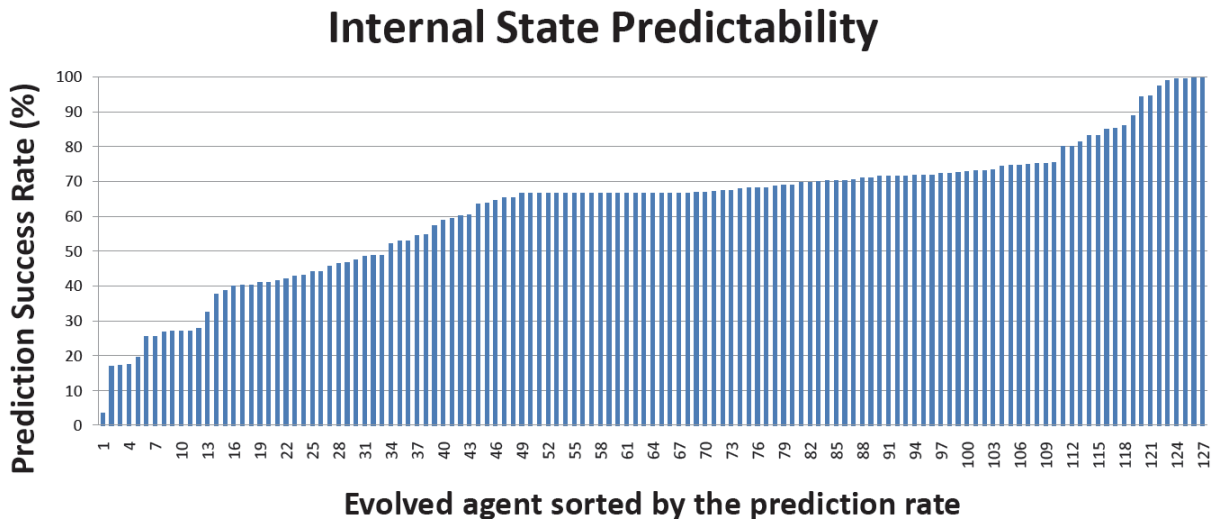


- Neural network predictor for a time series.

Method: Experimental Setup

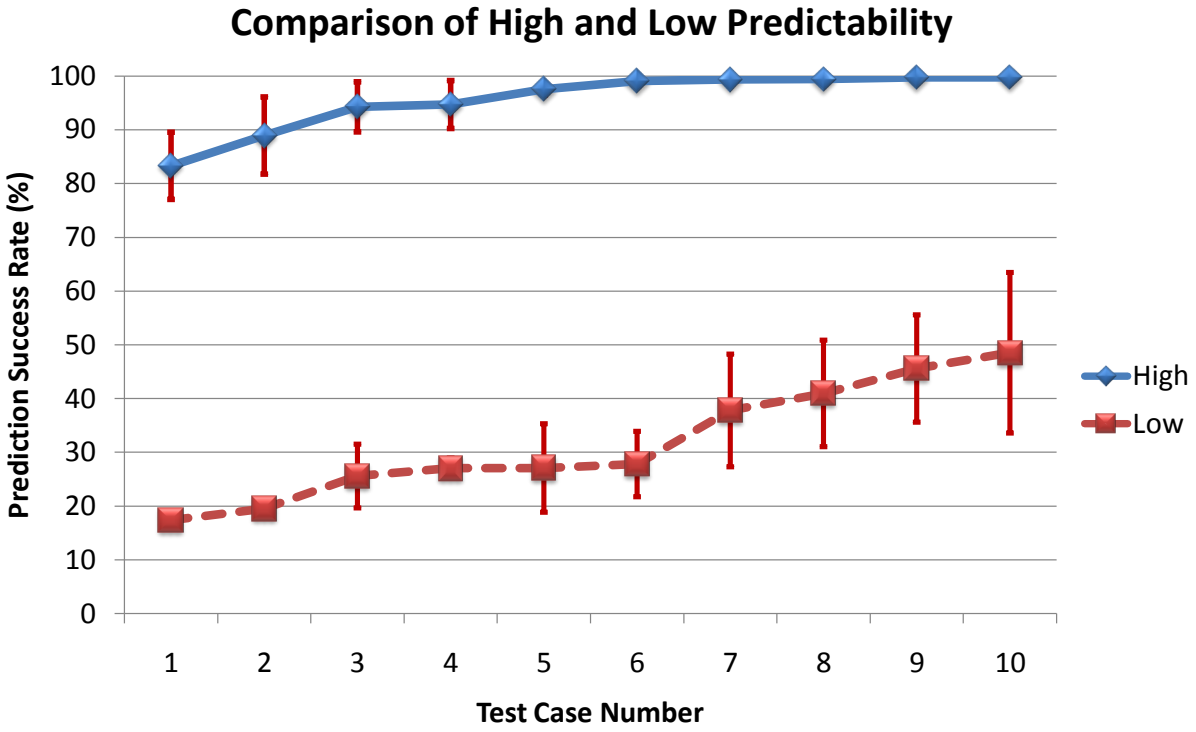
- Neural network predictor:
  - 2,000 training data.
  - 1,000 test data.
  - Back-propagation (learning rate 0.2).

Results: Internal State Predictability (ISP)



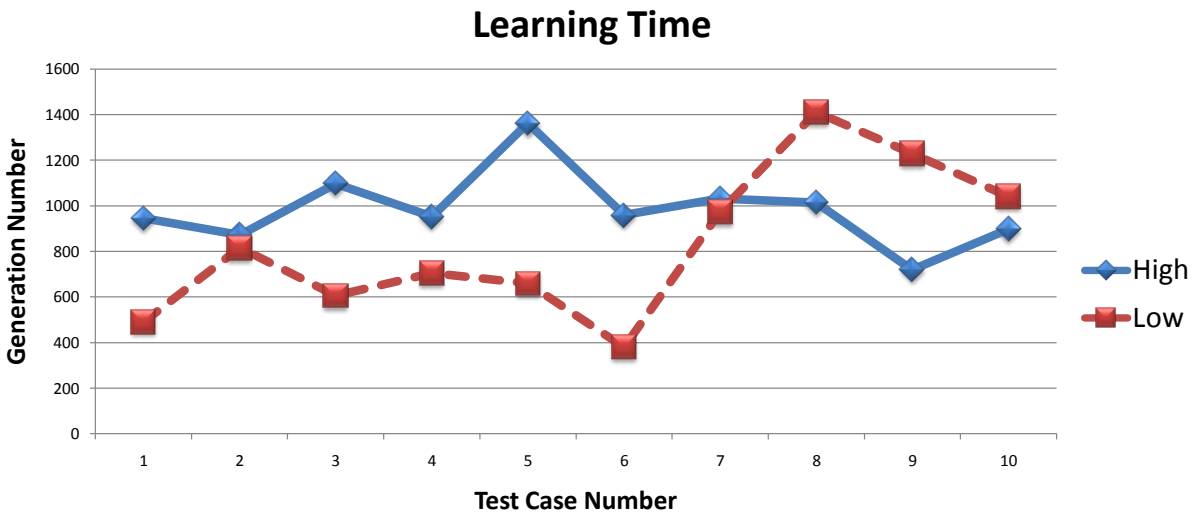
- Trained 130 pole balancing agents.
- Chose top 10 highest ISP agents and bottom 10 lowest ISP.
  - high ISPs: $\mu = 95.61\%$ and $\sigma = 5.55\%$.
  - low ISPs: $\mu = 31.74\%$ and $\sigma = 10.79\%$.

Comparison High ISP and Low ISP



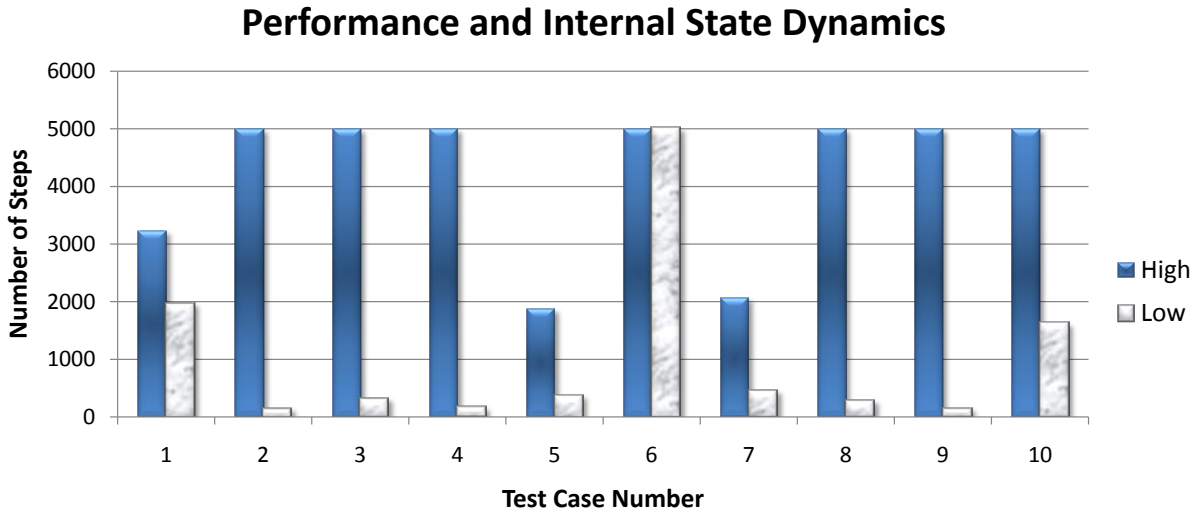
- A comparison of the average predictability from two groups: high ISP and low ISP.
- The predictive success rate of the top 10 and the bottom 10 agents.

Results: Learning Time



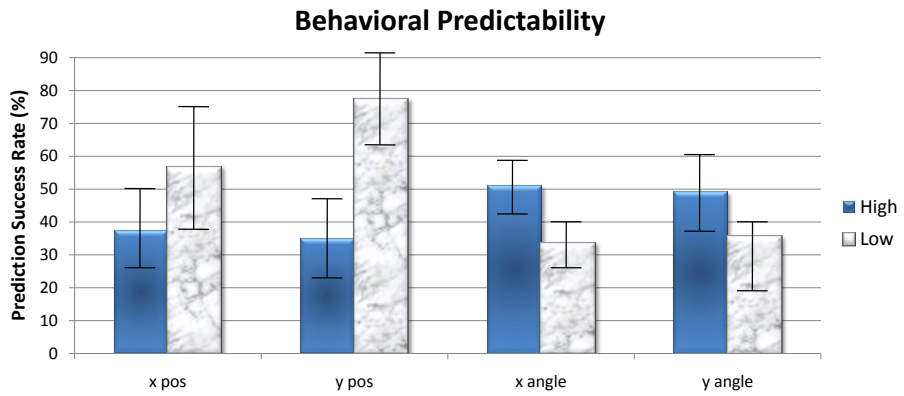
- No significant difference in learning time

Performance and Int. State Dyn.



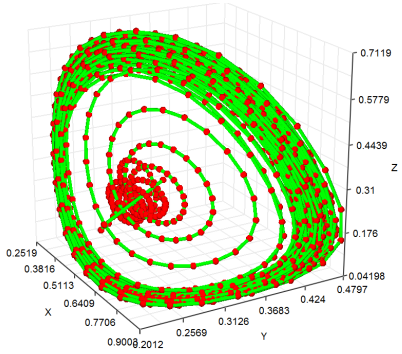
- Made the initial conditions in the 2D pole balancing task harsher.
- Performance of high- and low-ISP groups compared.
- High-ISP group outperforms the low-ISP group in the changed environment.

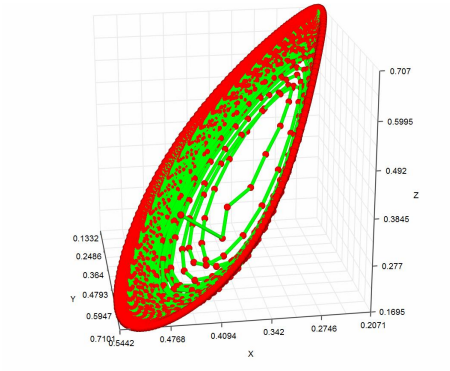
Behavioral Predictability

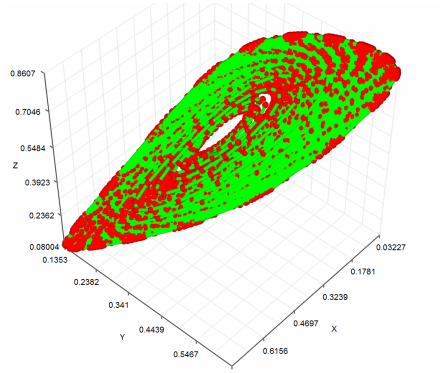


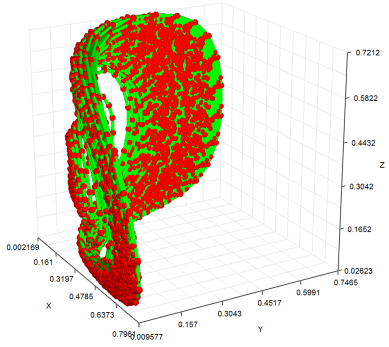
- Success of high-ISP group may simply be due to simpler behavioral trajectory.
- However, predictability in behavioral predictability is no different between high- and low-ISP groups.

Examples of internal state dynamics from the high ISP



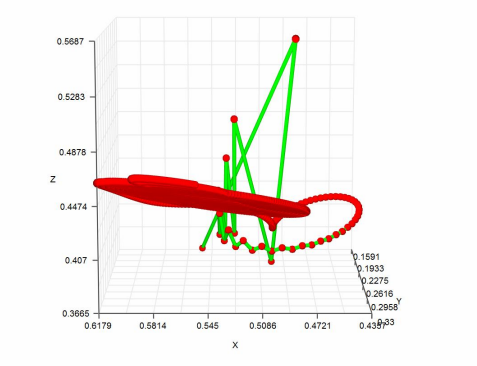


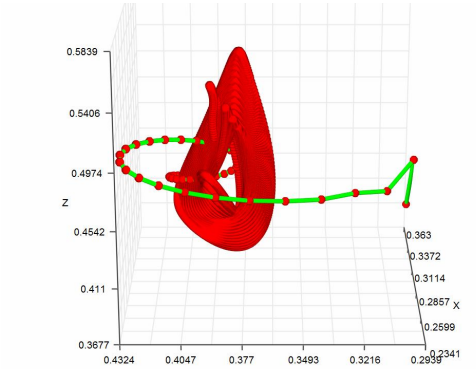


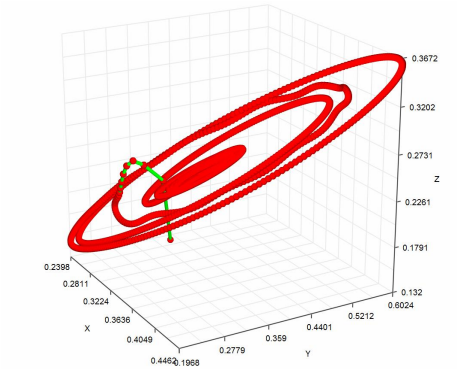


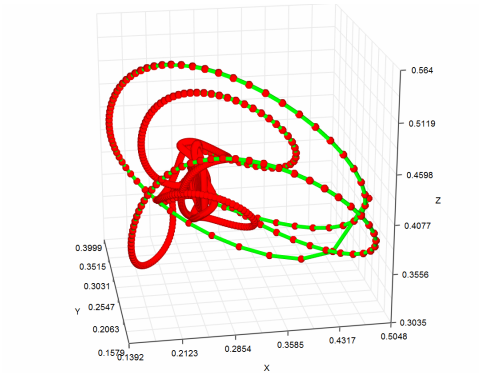
- Internal state dynamics show smooth trajectories.

Examples of internal state dynamics from the low ISP



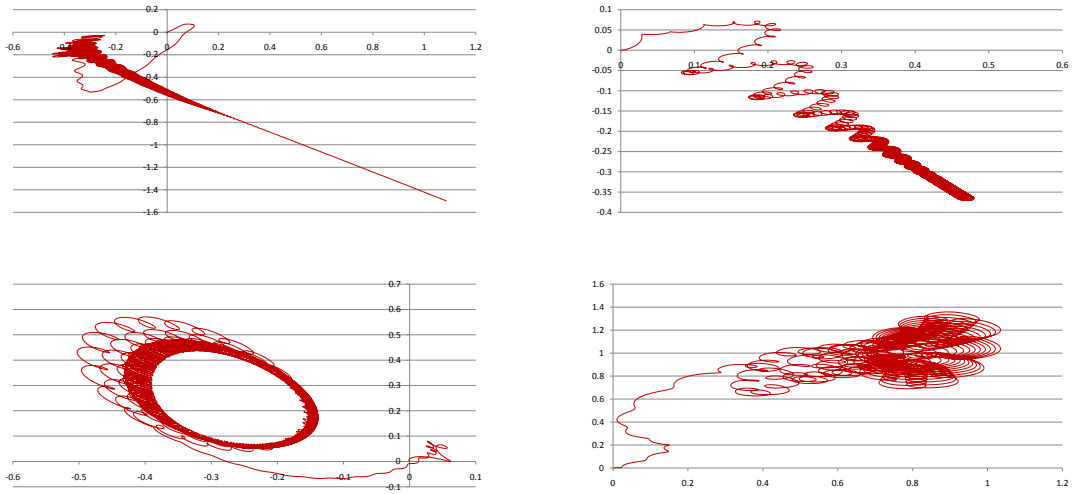






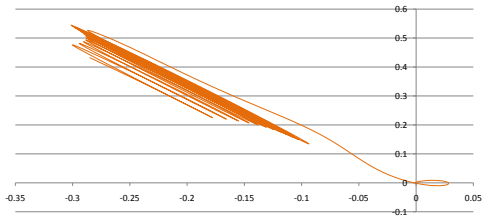
- Internal state dynamics show abrupt and jittery trajectories.

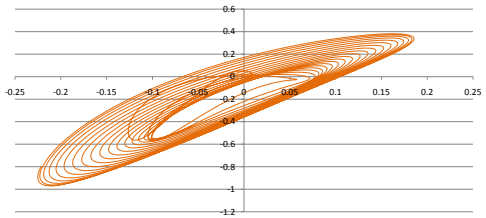
Examples of cart x and y position from high ISP

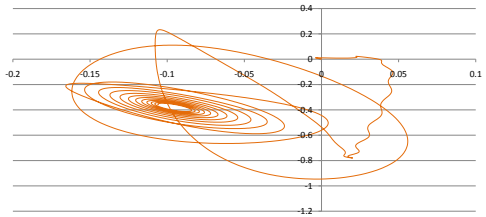


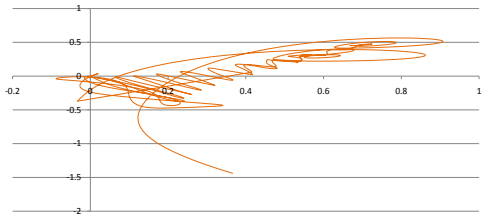
- Behavioral trajectories of x and y positions show complex trajectories.

Examples of cart x and y position from low ISP









- Behavioral trajectories of x and y positions show complex trajectories.

Related Work

- Bayesian self-model (Gold and Scassellati 2007).
- Continuous self re-modeling for resilient machines (Bongard et al. 2006).
- Autonomous mental development (Weng et al. 2001; Han et al. 2002).
- Role of self-awareness in cognition (Block 1995).
- Emergence of self-awareness from self-representation (Menant 2007).

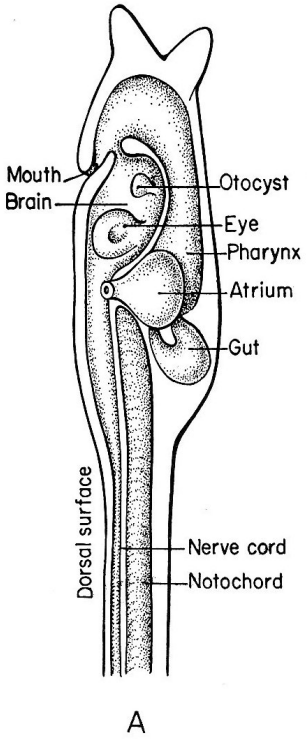
Conclusions

- Simpler (more predictable) internal dynamics can achieve higher levels of performance in harsher environmental conditions.
- The increased survival value is not always due to smoother behavior resulting from the simpler internal states.
- Initially evloution-transparent internal agent properties can affect external behavioral performance and fitness in a changing environment.
- An initial stepping stone in the evolutionary pathway leading to self-awareness and agency could have formed in such a way.

Why Do We Have a Brain?

Tree

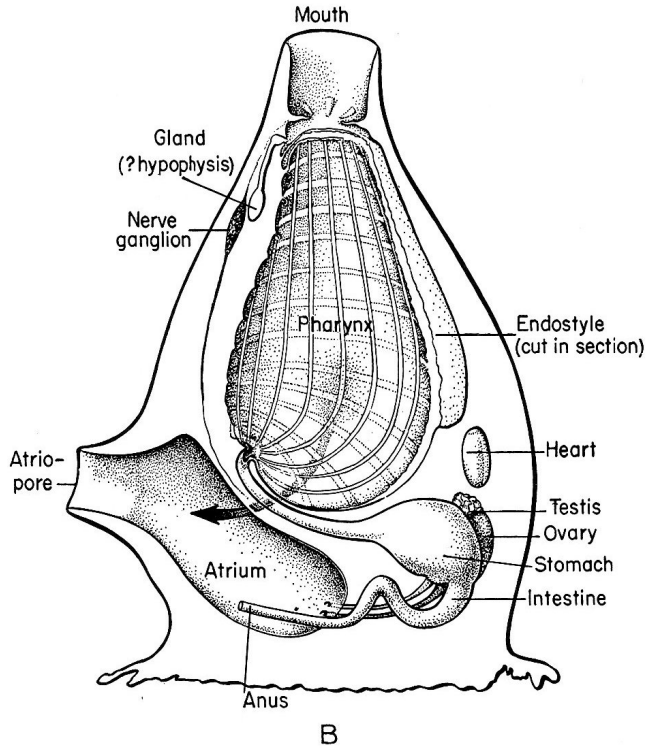
(no Brain)



Tunicate

Free-floating

(w/ Brain)



Tunicate

Settled

(w/o Brain)

- Brain vs. no brain (cf. Llinás et al. 1994).

Sources: http://homepages.inf.ed.ac.uk/jbednar/ and http://bill.srnr.arizona.edu/classes/182/Lecture-9.htm

References

Block, N. (1995). On a confusion about a function of consciousness. In *Behavoral and Brain Sciences*, 227–247.

Bongard, J., Zykov, V., and Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314:1118–1121.

Gold, K., and Scassellati, B. (2007). A Bayesian robot that distinguishes ”self” from ”other.”. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*. Nashville, Tennessee.

Han, J. D., Zeng, S. Q., Tham, K. Y., Badgero, M., and Weng, J. Y. (2002). Dav: A humanoid robot platform for autonomous mental development. In *Proceedings of the 2nd International Conference on Development and Learning*, 73–81. Cambridge, Massachusetts.

Llinás, R., Ribary, E., Joliot, M., and Wang, G. (1994). Content and context in temporal thalamocortical binding. In Buzsaki, G., editor, *Temporal Coding in the Brain*. Berlin: Springer Verlag.

Menant, M. C. (2007). Proposal for an approach to artificial consciousness based on self-consciousness.

Nolfi, S., Elman, J. L., and Parisi, D. (1994). Learning and evolution in neural networks. *Adaptive Behavior*, 3:5–28.

Weng, J., McClelland, J. L., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291(5504):599–600.