

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

SfN 2008 (738.14)

November 19, 2008

Jaerock Kwon and **Yoonsuck Choe**

Department of Computer Science

Texas A&M University

For the full paper, see Kwon and Choe (2008).

Abstract

What is the evolutionary value of self-awareness and agency in intelligent agents? One way to make this problem tractable is to think about the necessary conditions that lay the foundation for the emergence of agency, and assess their evolutionary origin. We postulate that one such requirement is the predictability of the internal state trajectory. A distinct property of one's own actions compared to someone else's is that one's own is highly predictable, and this gives the sense of "authorship". In order to investigate if internal state predictability has any evolutionary value, we evolved sensorimotor control agents driven by a recurrent neural network in a 2D pole-balancing task. The hidden layer activity of the network was viewed as the internal state of an agent, and the predictability of its trajectory was measured. We took agents exhibiting equal levels of performance during evolutionary trials, and grouped them into those with high or low internal state predictability (ISP). The high-ISP group showed better performance than the low-ISP group in novel tasks with substantially harder initial conditions. These results indicate that regularity or predictability of neural activity in internal dynamics of agents can have a positive impact on fitness, and, in turn, can help us better understand the evolutionary role of self-awareness and agency.

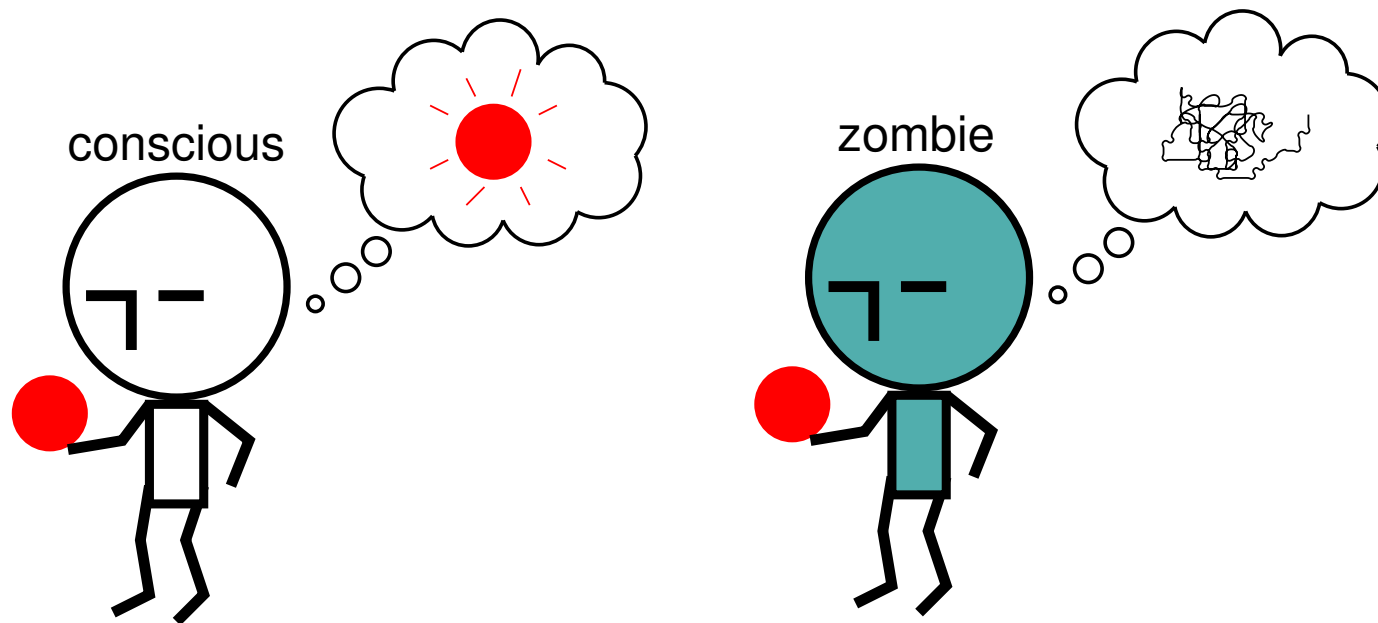
Research Question: Self-Awareness

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.

Evolution of Self-Awareness and Agency?



- Performance-wise, **conscious agents** and **zombies** could be indistinguishable (to evolution)!

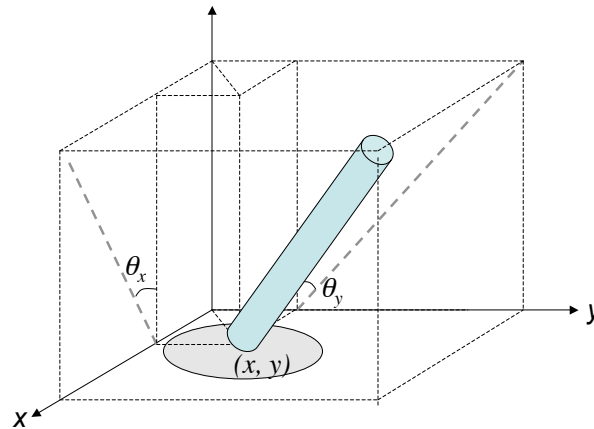
Approach

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.

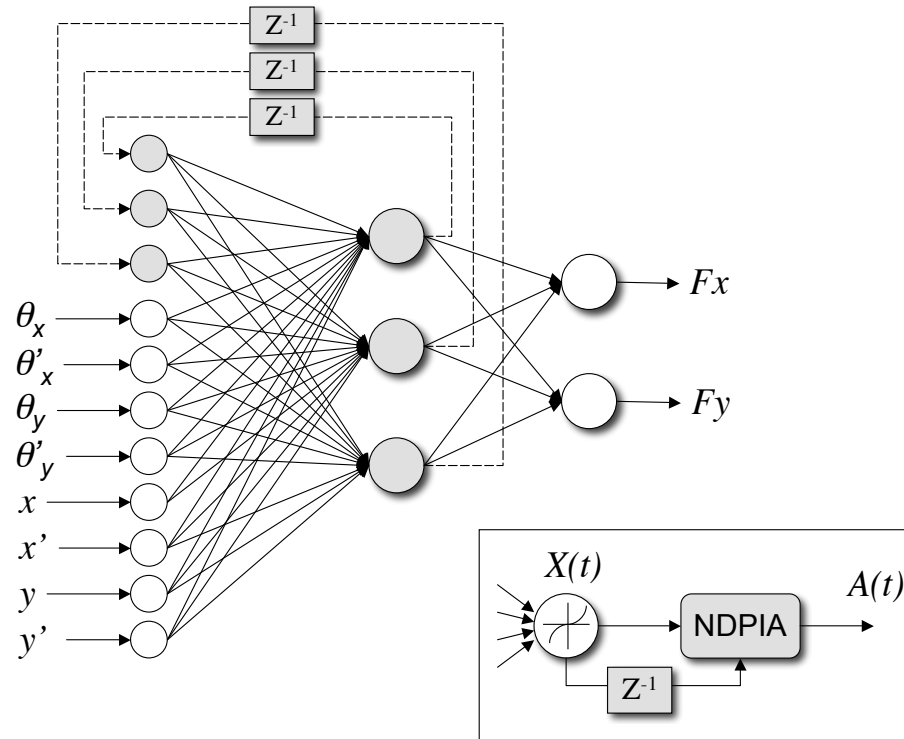
Necessary condition identified: Need to be able to **predict** one’s own internal state (cf. Nolfi et al. 1994).

Method (Task): 2D Pole-Balancing



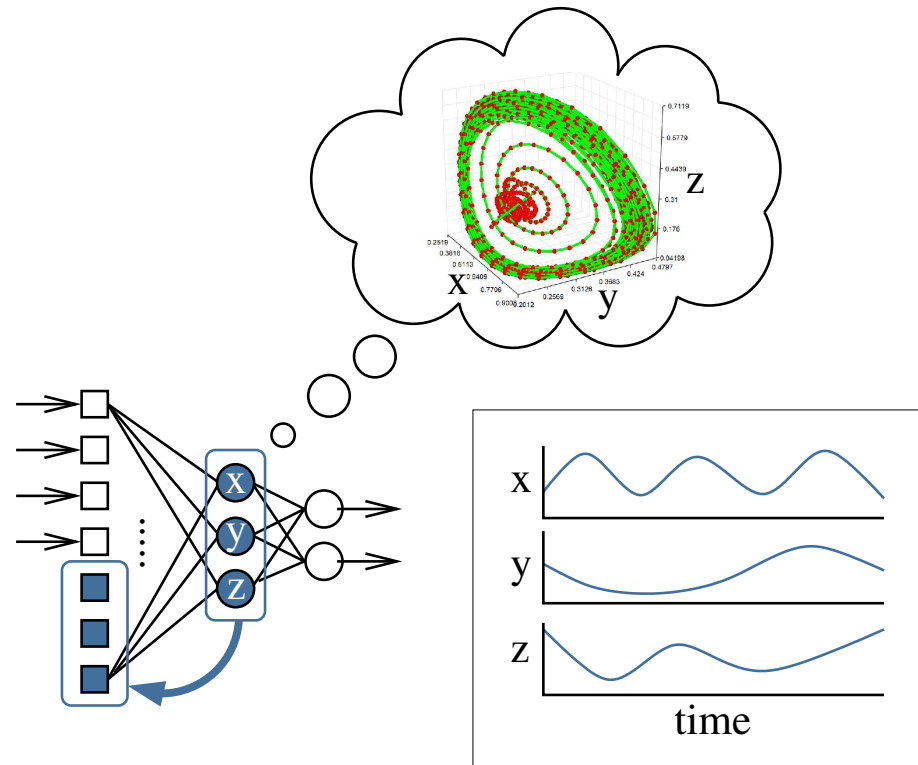
- Physical parameters of the pole balancing system: position (x, y) ; velocity (\dot{x}, \dot{y}) ; pole angle (θ_x, θ_y) ; angular velocity $(\dot{\theta}_x, \dot{\theta}_y)$.

Method: Neuroevolution Controller



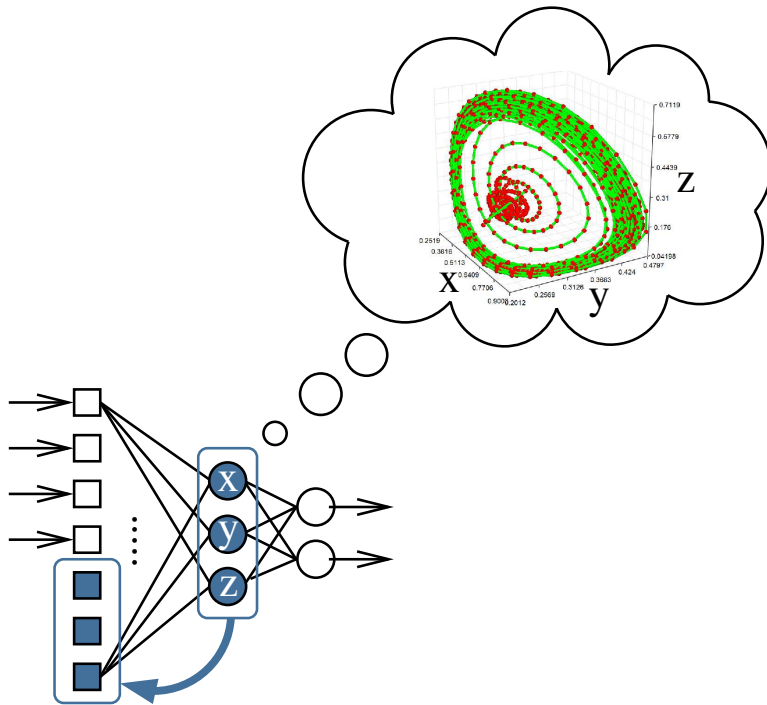
- Recurrent neural network for 2D pole balancing.
- Trained with standard neuroevolution.
- Investigate the internal state trajectories.

Internal State of the Controller

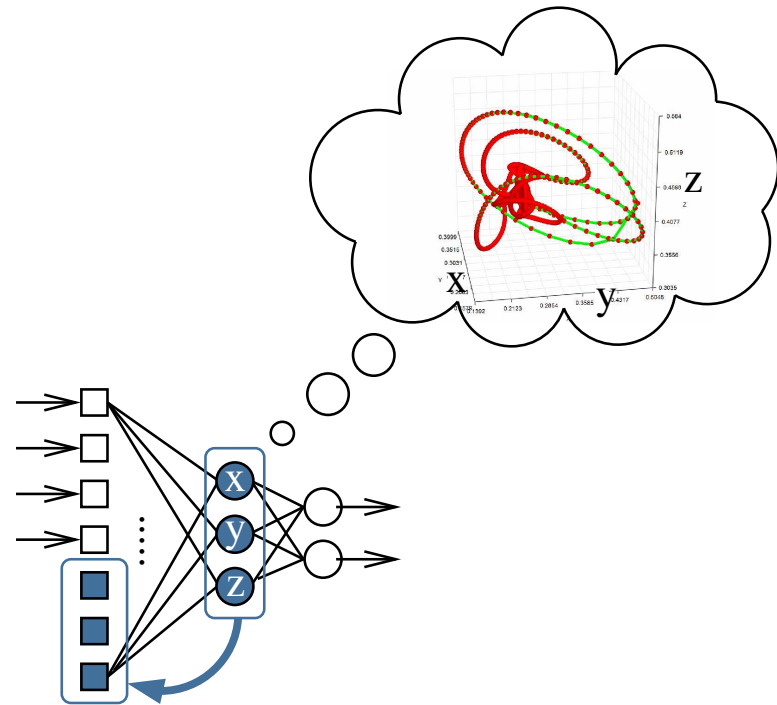


- Activation level of hidden units can be seen as the **internal state** of the controller agent.

Same Behavior, Different Mind



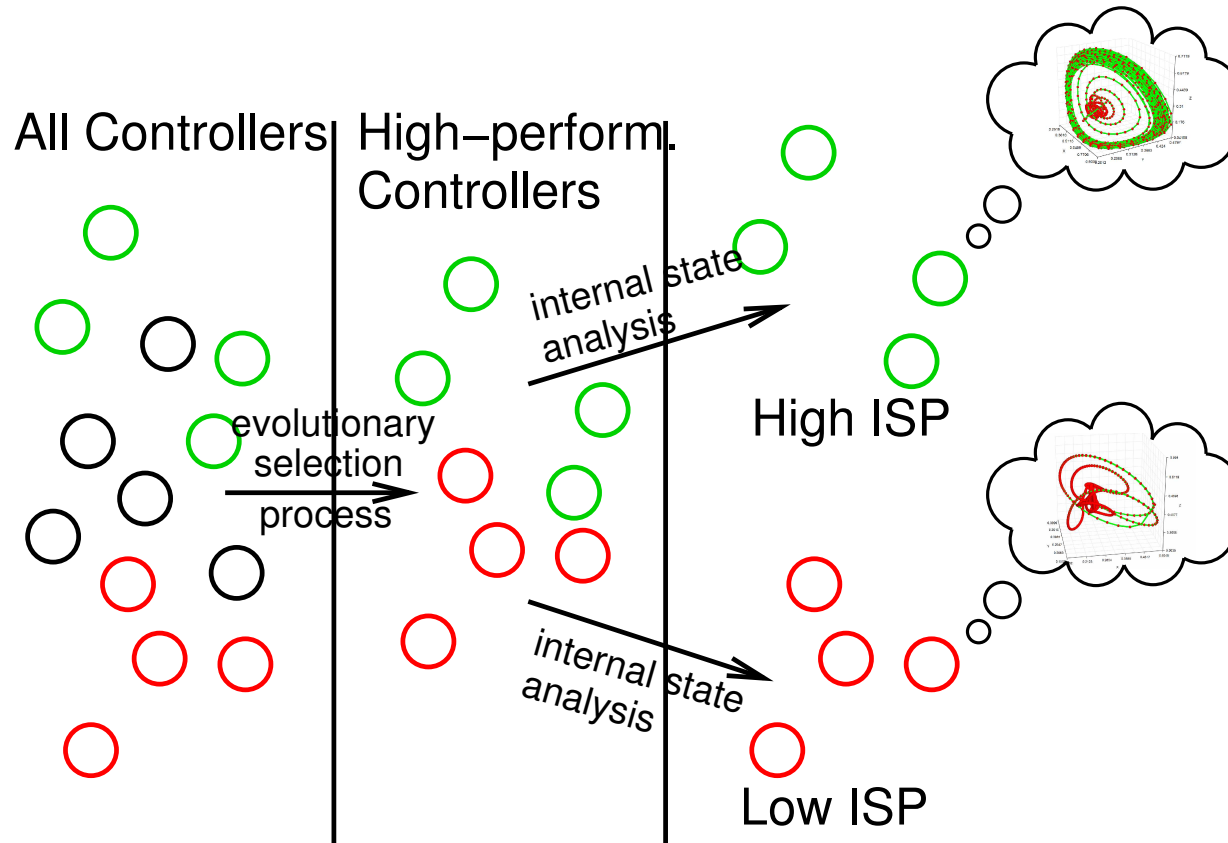
(a) High Internal State Predictability



(a) Low Internal State Predictability

- Two controllers with the same level of performance can have different internal state dynamics!

Sketch of the Method

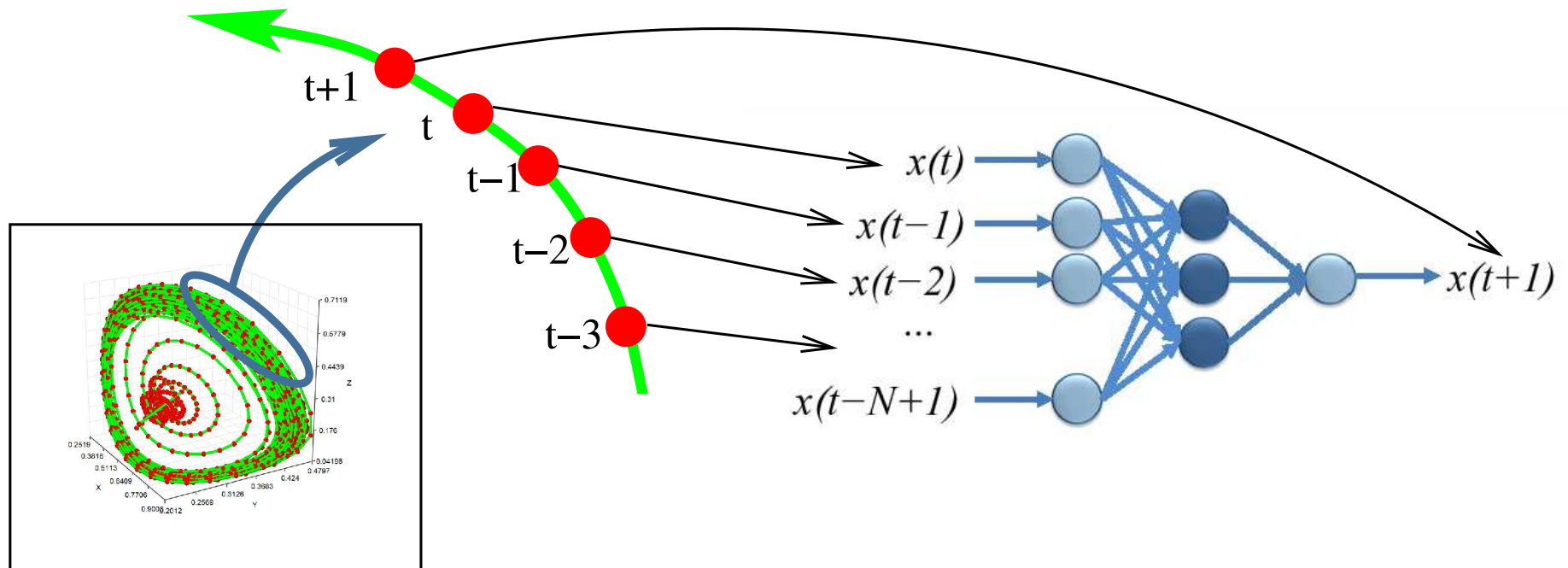


1. Evolve controllers to meet a fixed performance criterion (fitness does not measure predictability) in pole-balancing tasks.
2. Group high-performance individuals into high- and low internal state predictability (ISP) groups.

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

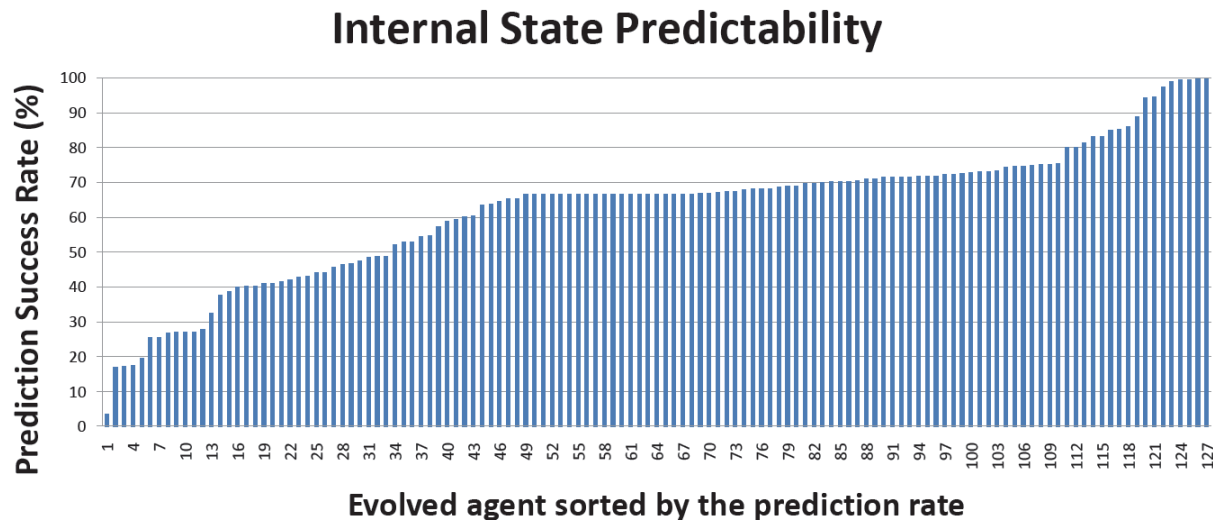


Neural network predictor (backprop):

- Input: hidden unit activation in N steps in the past
- Target: current hidden unit activation

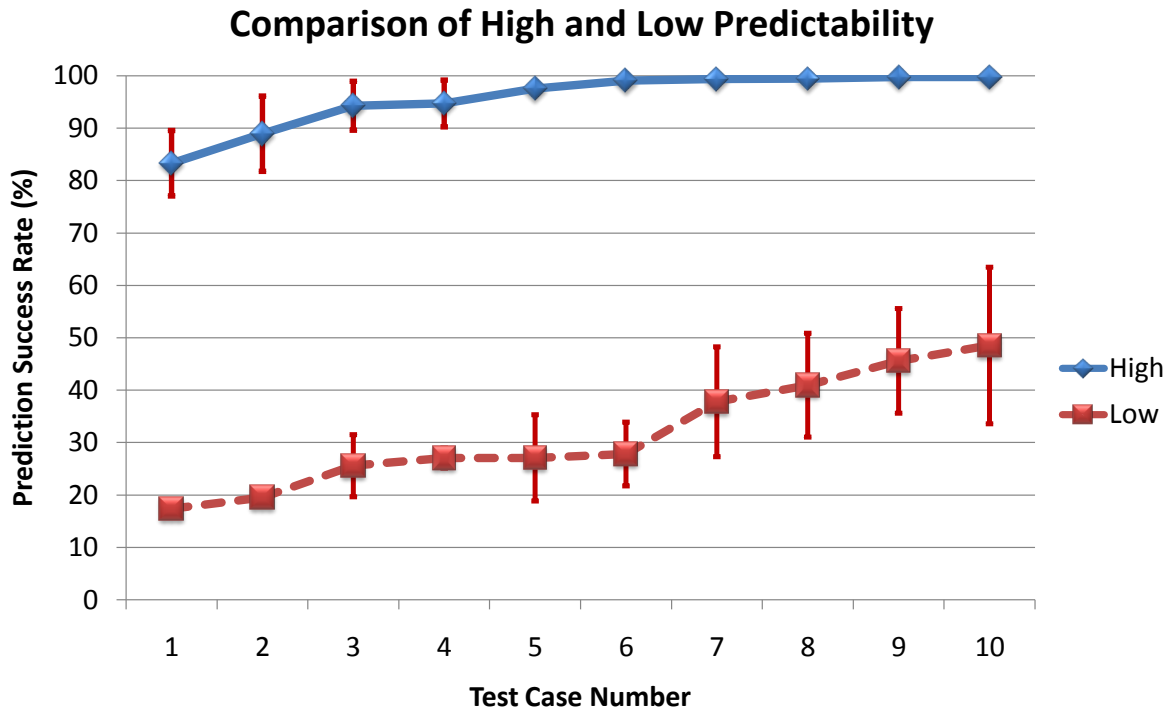
Measure how easy it is to learn to predict trajectory.

Results: Internal State Predictability (ISP)



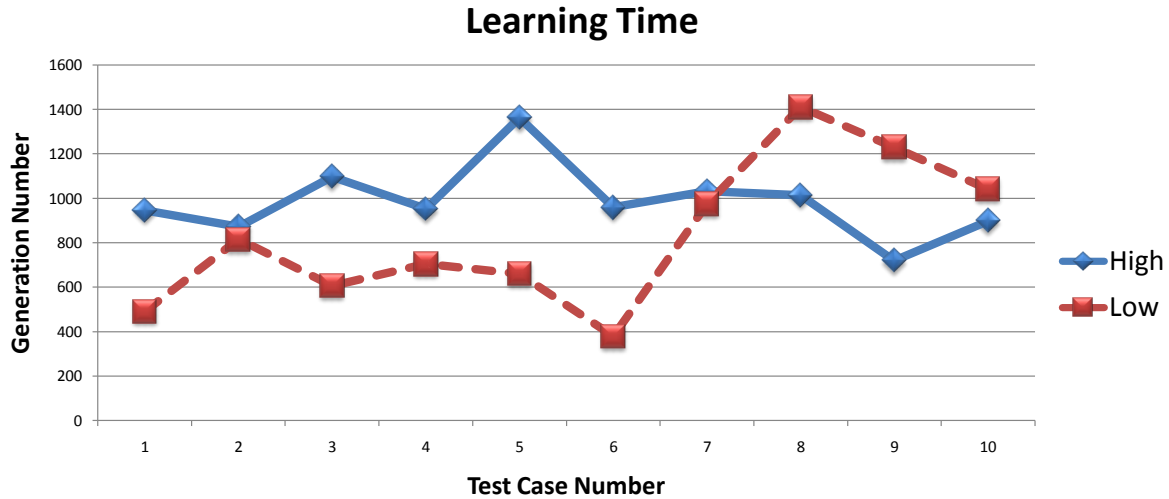
- Trained 130 pole balancing agents.
- Chose top 10 highest ISP agents and bottom 10 lowest ISP.
 - high ISPs: $\mu = 95.61\%$ and $\sigma = 5.55\%$.
 - low ISPs: $\mu = 31.74\%$ and $\sigma = 10.79\%$.

Comparison High ISP and Low ISP



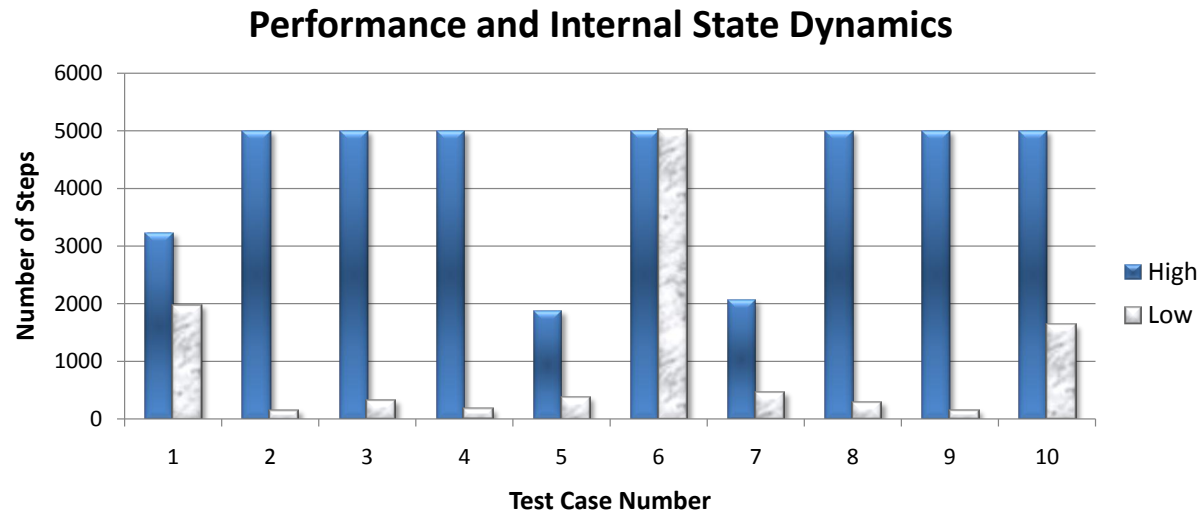
- A comparison of the average predictability from two groups: high ISP and low ISP.
- The predictive success rate of the top 10 and the bottom 10 agents.

Results: Learning Time



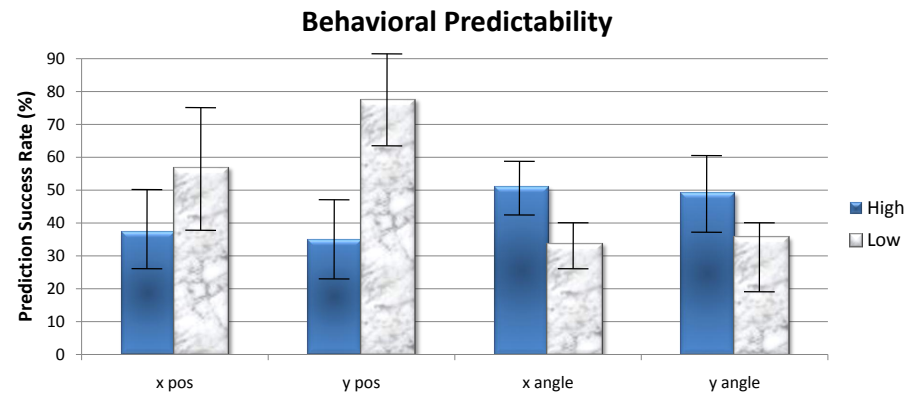
- No significant difference in learning time

Performance and Int. State Dyn.



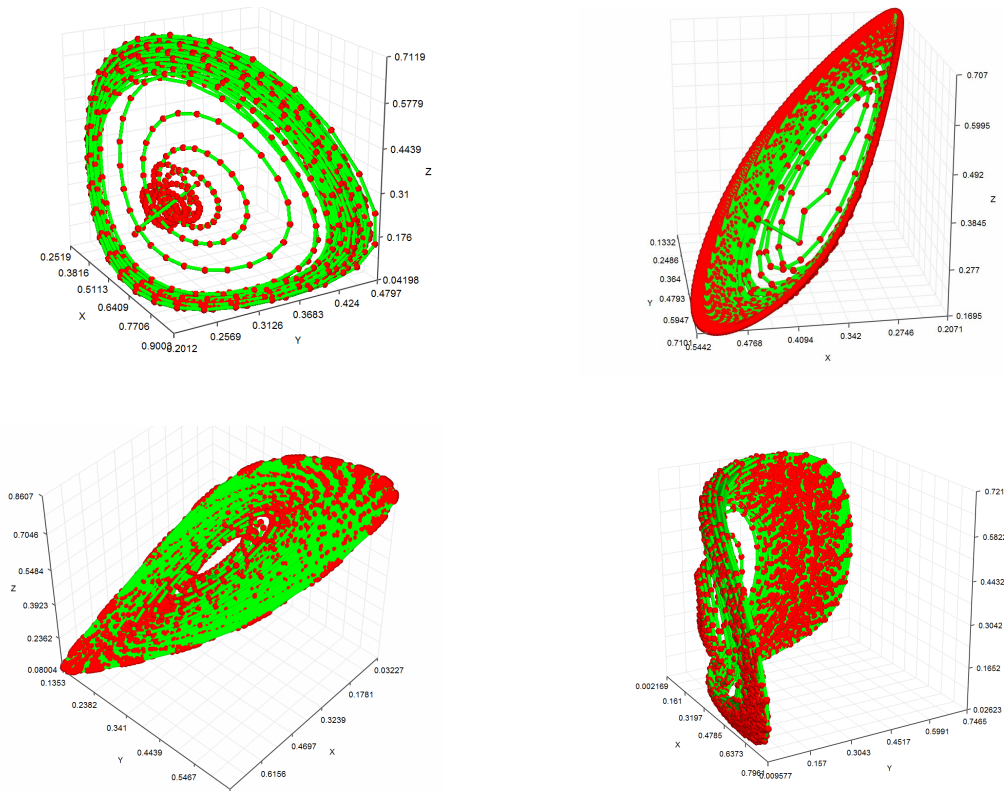
- Made the initial conditions in the 2D pole balancing task harsher.
- Performance of high- and low-ISP groups compared.
- High-ISP group outperforms the low-ISP group in the changed environment.

Behavioral Predictability



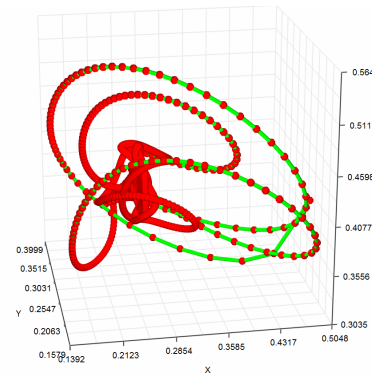
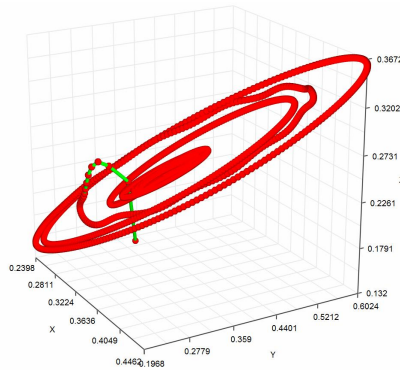
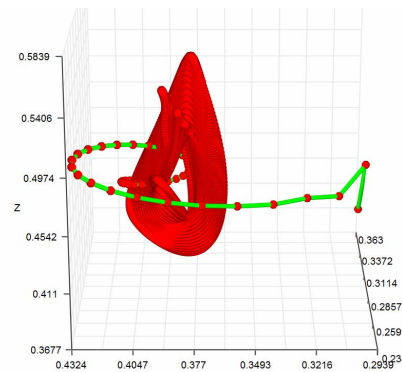
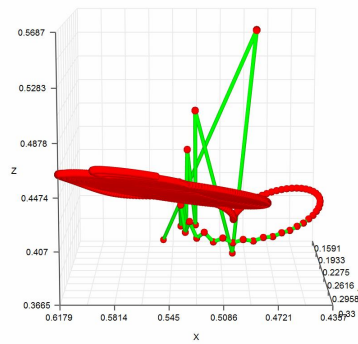
- Success of high-ISP group may simply be due to simpler behavioral trajectory.
- However, predictability in behavioral predictability is no different between high- and low-ISP groups.

Examples of Internal State Dynamics from the High ISP Group



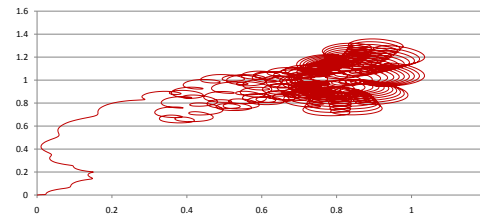
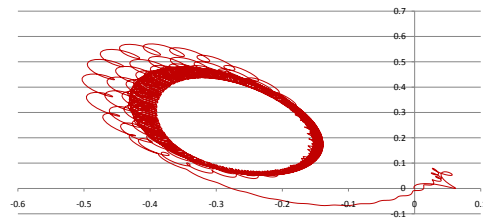
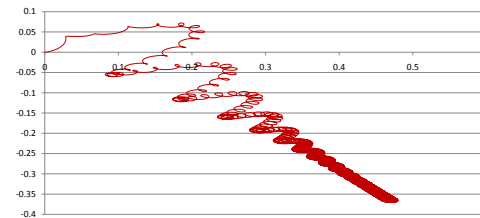
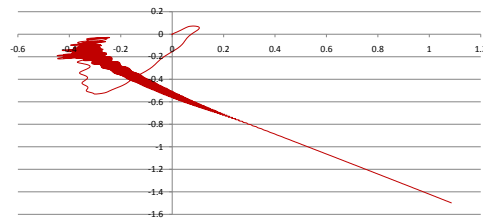
- Internal state dynamics show smooth trajectories.

Examples of Internal State Dynamics from the Low ISP Group



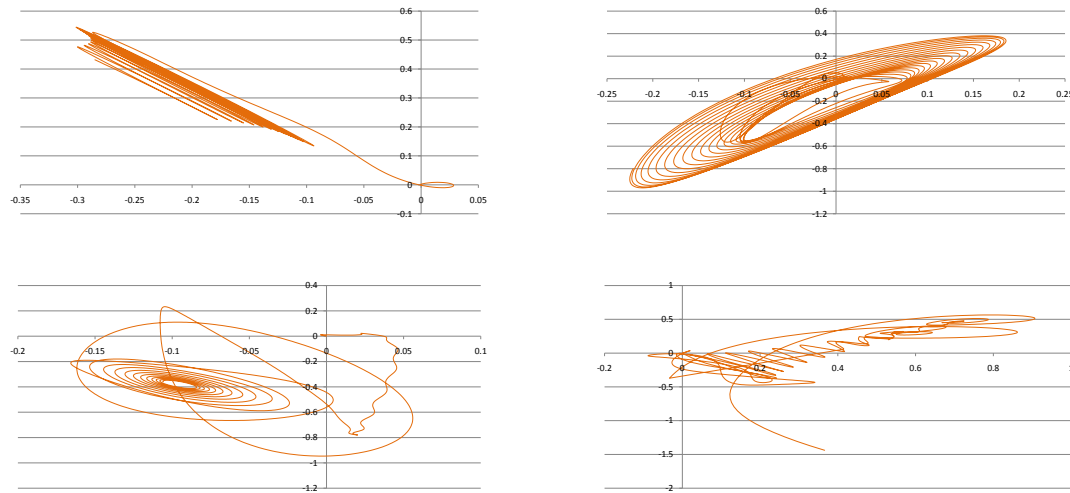
- Internal state dynamics show abrupt and jittery trajectories.

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 evolution-transparent internal agent properties can affect external behavioral performance and fitness in a changing environment.
- Speculation: Maybe this is how self-awareness/agency evolved?

References

Block, N. (1995). On a confusion about a function of consciousness. In *Behavioral 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.

Kwon, J., and Choe, Y. (2008). Internal state predictability as an evolutionary precursor of self-awareness and agency. In *Proceedings of the Seventh International Conference on Development and Learning*.

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.