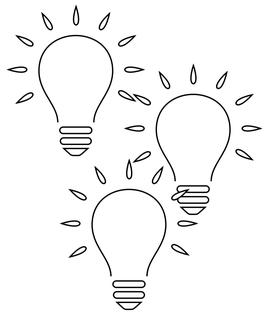
Investigating Curious Behaviour in Reinforcement Learning Agents

Nadia M. Ady and Patrick M. Pilarski

Dept. of Computing Science & Dept. of Medicine, University of Alberta, Edmonton, AB, Canada

Researchers have been **curious** about computational curiosity for a long time. Many researchers have suggested different **learning signals** to motivate curiosity.

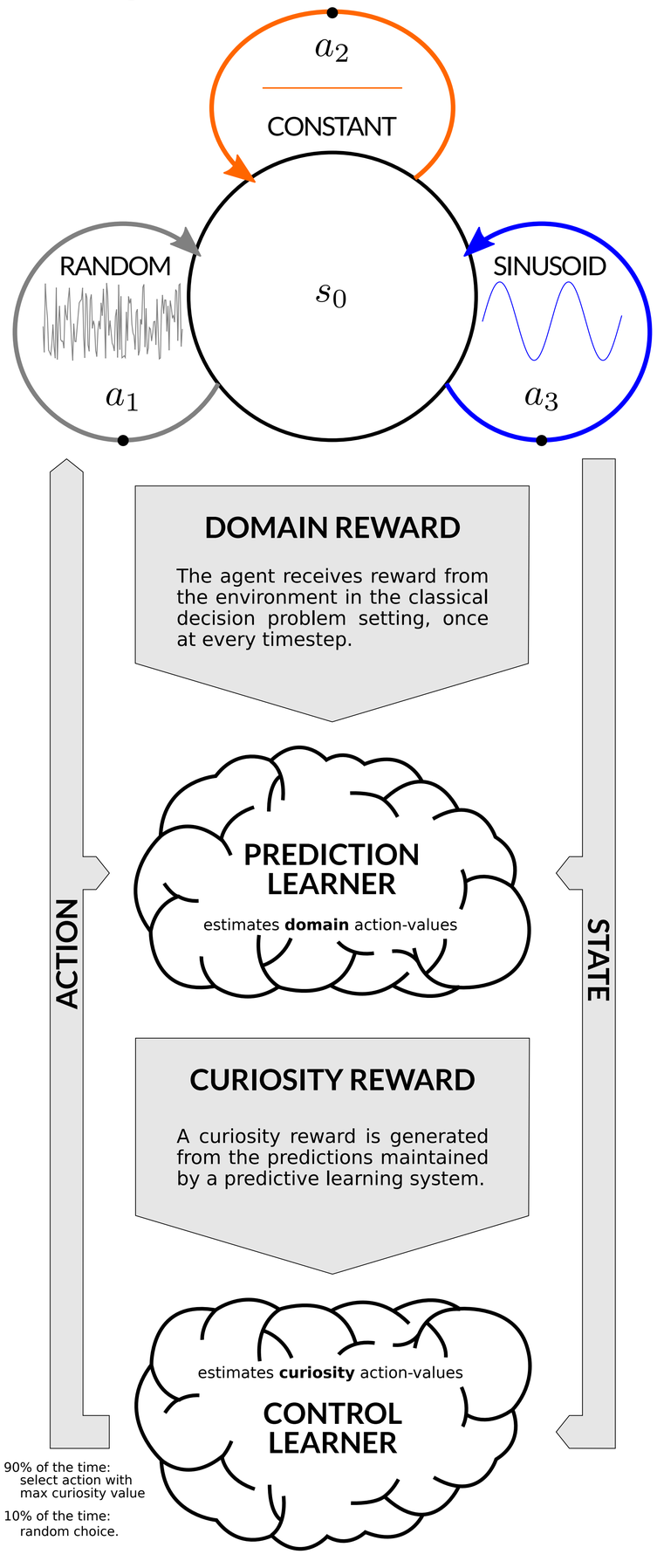


But how can we compare and understand the differences in **behaviour** stemming from different signals for this purpose?

Small domains allow us to see differences in behaviour between different methods for curiosity in reinforcement learning agents, often hidden by more complex domains

THE CURIOSITY BANDIT

is designed to showcase **variation in domain reward**

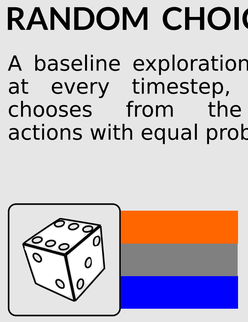


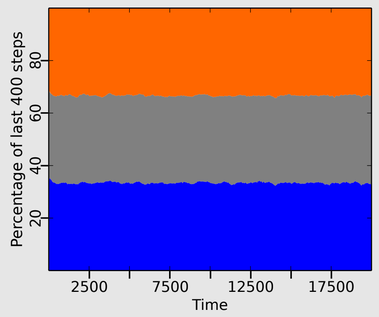
This work was supported by the Alberta Machine Intelligence Institute (AMII), Alberta Innovates Technology Futures (AITF), and the Natural Sciences and Engineering Research Council (NSERC).

PROPOSED CURIOSITY REWARD SIGNALS

RANDOM CHOICE

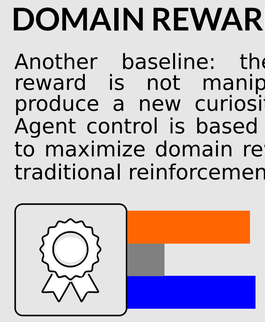
A baseline exploration algorithm, at every timestep, the agent chooses from the available actions with equal probability.

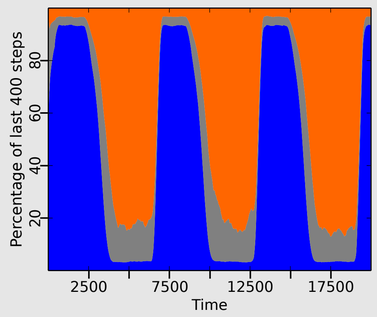




DOMAIN REWARD

Another baseline: the domain reward is not manipulated to produce a new curiosity reward. Agent control is based on aiming to maximize domain reward as in traditional reinforcement learning.



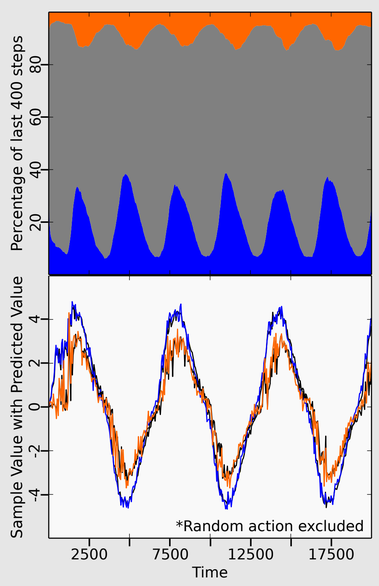


ABSOLUTE ERROR

Absolute prediction error was one of the earliest learning signals optimized specifically for curiosity (Schmidhuber 1991, in Meyer & Wilson (eds.) 222).

The initial intuition: to improve prediction, spend more time in areas of high error. Unfortunately, such an agent could get stuck in areas of high randomness.

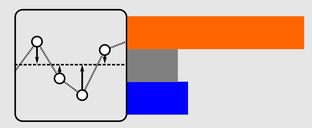
We see the predicted attraction to randomness and periodic spikes of the sinusoidal action exactly at domain value crests and troughs, where the estimate lags most.

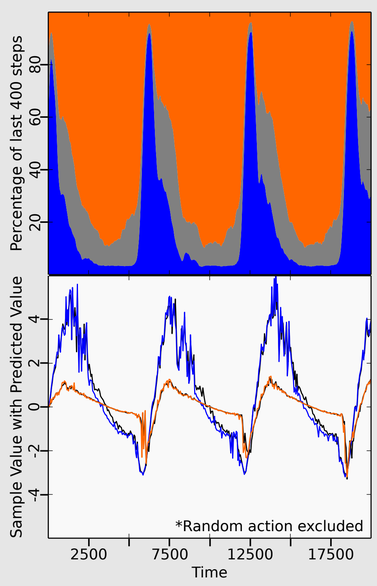


ERROR

Maximizing true-valued error in domain value, as used by Schembri et al. (2007, in Adv. in Artificial Life. P. of the 9th Eur. C. on Artificial Life, v. 4648, 294), is intuitively beneficial: an agent will favour areas which seem to be better than expected and to avoid areas which seem to be getting worse.

As occurred with absolute error, the agent gets the greatest error choosing the sinusoidal action at its crests, but at its troughs, the constant action is under-estimated, giving the best error.

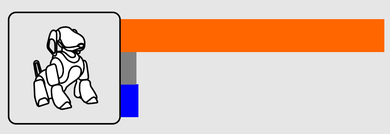


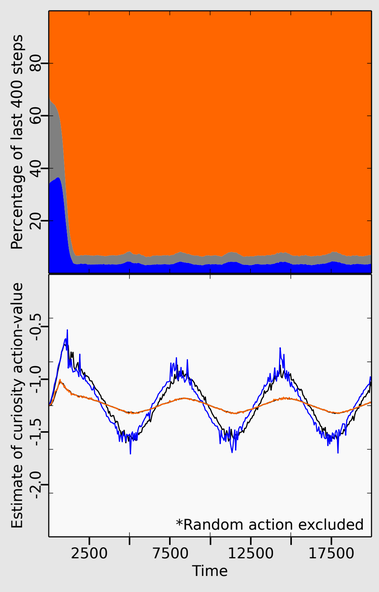


LEARNING PROGRESS

Oudeyer et al. (2007, in IEEE T. on Evolutionary Computation 11(2), 265) utilized a measure of learning progress, or the decrease in error. Intuitively, an agent achieves high learning progress with prediction improvement.

In taking the constant action, the agent shifts its value closer to zero by increasingly smaller amounts. The sinusoidal action only ever shows more learning progress at the crests and troughs, where the predicted and sample values cross, resulting in corresponding blips.



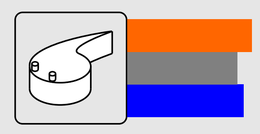


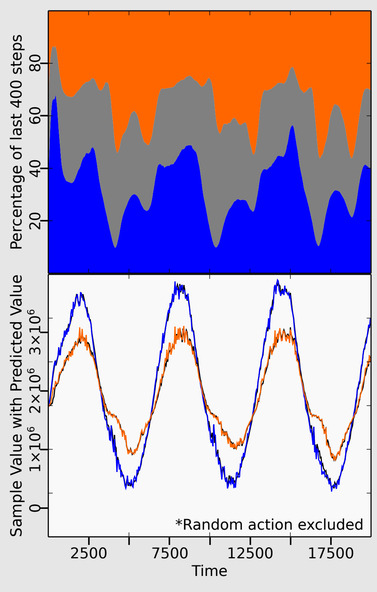
WHITE'S SURPRISE

White et al. (2014, in *Workshops at the AAAI C. on AI*) suggested a measure of surprise (also called unexpected demon error) which can be used as a curiosity reward.

White's surprise is the ratio of the moving average of the error and the variance of the error.

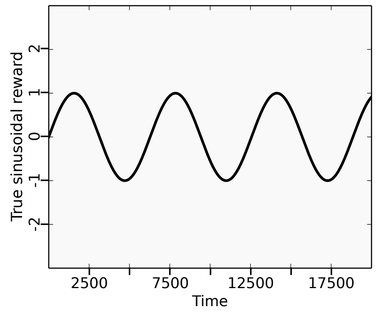
We see two peaks of the constant action either side of each trough and two peaks of the sinusoidal action either side of each crest.





Compare this plot to those above.

We can see how the variations in the domain reward impact the behaviour of agents optimizing for different curiosity rewards.



SO WHAT?

It is unclear what the "right" system for curiosity should be. But it is important to recognize how they differ to build off a rich historical foundation of ideas.

UP NEXT...

CURIOSITY RING WORLD

designed to showcase **variation in state-action dynamics**

RANDOM

CONSTANT

PERIODIC