# Curiosity-Driven Exploration with Planning Trajectories

**Tyler Streeter**

Virtual Reality Applications Center, Iowa State University
1620 Howe Hall, Ames, IA 50010
http://www.vrac.iastate.edu/~streeter
tylerstreeter@gmail.com

## Introduction

Reinforcement learning (RL) agents can reduce learning time dramatically by planning with learned predictive models. Such planning agents learn to improve their actions using planning trajectories, sequences of imagined interactions with the environment. However, planning agents are not intrinsically driven to improve their predictive models, which is a necessity in complex environments. This problem can be solved by adding a curiosity drive that rewards agents for experiencing novel states. Curiosity acts as a higher form of exploration than simple random action selection schemes because it encourages targeted investigation of interesting situations.

In a task with multiple external rewards, we show that RL agents using uncertainty-limited planning trajectories and intrinsic curiosity rewards outperform non-curious planning agents. The results show that curiosity helps drive planning agents to improve their predictive models by exploring uncertain territory. To the author's knowledge, no previous work has tested the benefits of curiosity with planning trajectories.

## Models of Curiosity

The model of curiosity used in Barto, Singh, & Chentanez (2004) rewards agents for experiencing novel states. The authors used this model in a RL framework that developed hierarchical sets of skills, driven by curiosity rewards.

Schmidhuber (1991) described a model of curiosity that rewards agents when prediction errors decrease over time. This method is more robust than simply rewarding agents in novel states. It avoids the problem of attraction to random signals in non-deterministic environments.

Oudeyer & Kaplan (2004) presented a mechanism called "Intelligent Adaptive Curiosity" (IAC) which drives agents to explore situations that are neither too predictable nor too unpredictable. This has similarities to Schmidhuber's (1991) model. However, IAC helps solve a subtle problem: agents sometimes learn to alternate between unpredictable and predictable situations, causing them to receive rewards for reducing their prediction uncertainty. To overcome this, IAC tracks similar situations and measures uncertainty reduction only within a specific situation.



## Agent Architecture

The architecture used here is designed as an actor-critic model with temporal difference learning. Observations are converted to a radial basis function state representation. The value function and policy are represented as simple linear neural networks. Planning is performed using a set of predictor neural networks that, given an observation and hypothetical action, output a predicted next observation and reward, along with a metaprediction of the agent's own prediction uncertainty. These predictors are trained with the actual next observation, reward, and prediction error, respectively.

At each time step, first the predictors output their predictions and are trained with the actual data received. Then an uncertainty-limited planning trajectory begins, starting from the current actual observation. This trajectory continues until either: 1) the maximum trajectory length is exceeded, or 2) the prediction uncertainty exceeds some threshold. At each step of the planning trajectory, the policy uses a softmax action selection scheme. The value function and policy are trained with temporal difference learning, with the total reward equal to the predicted current reward plus a value proportional to the prediction uncertainty (i.e. the curiosity reward). Thus, the curiosity model used here is most similar to that of Barto, Singh, & Chentanez (2004). (This relatively simple model is sufficient for the deterministic task used in the next section.) Note that reinforcement learning is performed using *only* predicted values; the actual values are used only to train the predictors.

Temporal difference learning, actor-critic models, and planning trajectories are described in Sutton & Barto (1998). For more details about the present architecture, including learning update equations and detailed diagrams, see Streeter (2005). An open source implementation of this architecture, images and videos of applications, etc. are available at: http://verve-agents.sourceforge.net and http://www.vrac.iastate.edu/~streeter/AAAI-2006.

## Results

To test the benefits of planning trajectories with curiosity, the discrete environment shown in Figure 1 is used. An agent in this environment is able to sense its 2-dimensional position and can move left, right, up, down, or remain in place. Three spaces in the environment contain small rewards, and one contains a large reward. Performance is measured as the sum of rewards received over the course of a single trial. Each trial lasts 100 time steps.

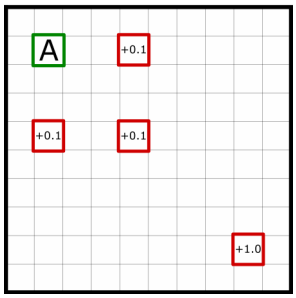


Figure 1: A discrete 2D world with multiple rewards. The space marked "A" is the agent's starting location. Spaces with positive numbers indicate reward locations. All other spaces yield rewards of zero.

The performance of three RL agents is compared: one agent without planning or curiosity, one with planning, and one with planning and curiosity. The hypothesis is that curious agents will be driven to explore the entire environment, eventually finding the larger reward. Non-curious agents, on the other hand, will quickly find the small rewards and will not be motivated to continue searching.

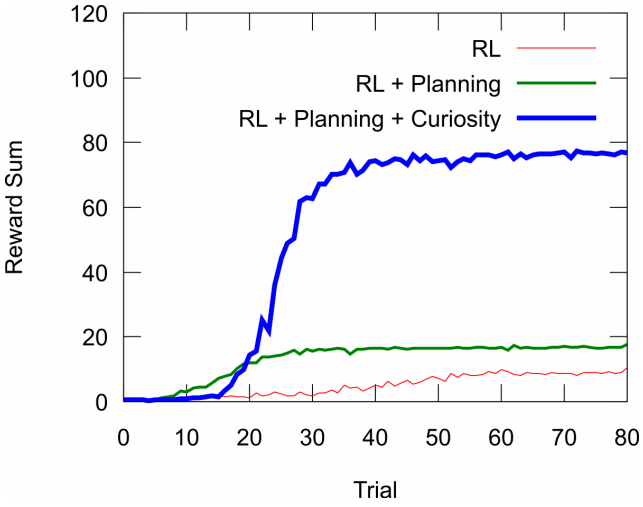


Figure 2: Learning performance of three RL agents: one without planning or curiosity, one with planning, and one with planning and curiosity. Results are averaged over 50 runs. Planning agents were allowed a maximum of 50 steps per planning trajectory.

The results in Figure 2 show the three agents' reward sums over time. The agent with planning outperforms the one without planning because it is able to train its value function and policy many times per step during a planning trajectory. Curiosity is clearly beneficial on this task, as evidenced by the higher sum achieved by the curious agent. The non-curious planning agent quickly finds one of the small rewards, but then it trains its value function and policy to focus on this treasure without worrying about the rest of the state space. The curious agent, with its intrinsic drive to experience novel states, consistently finds the larger goal, yielding a higher reward sum.

The curious agent plot shows a noticeable delay before any visible improvement. This is a curiosity-driven exploration phase. During this time the curious agent is mainly driven by curiosity rewards. Around trial 15, its prediction uncertainty is low, and its predictive model is accurate, so it no longer generates intrinsic curiosity rewards. Instead, it begins maximizing external reward intake.

## Discussion

We have shown, using a reinforcement learning task, that curiosity is beneficial to agents using planning trajectories. One of the main purposes of planning is to reduce the number of trials needed to learn a task (i.e. to train a value function and policy). Curiosity drives the agent to improve its predictive model, increasing the overall effectiveness of planning. In complex environments with multiple external rewards, curiosity is essential. It promotes targeted exploration towards uncertain territory.

## References

Schmidhuber, J. 1991. Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks*, vol. 2, 1458-1463. IEEE.

Barto, A., Singh, S., & Chentanez, N. 2004. Intrinsically motivated learning of hierarchical collections of skills. In *3rd International Conference on Development and Learning.*

Oudeyer, P.-Y., and Kaplan, F. 2004. Intelligent adaptive curiosity: a source of self-development. In Berthouze, L., et al, eds., *Proceedings of the 4th International Workshop on Epigenetic Robotics*, volume 117, 127-130. Lund University Cognitive Studies.

Streeter, T. 2005. Design and implementation of general purpose reinforcement learning agents. Unpublished master's thesis, http://www.vrac.iastate.edu/~streeter.

Sutton, R., and Barto, A. 1998. *Reinforcement learning: an introduction*. MIT Press.