Predictive Improvement through Latent Space Optimisation

Extended Abstract

Alexander McCaffrey
City, University of London
London, United Kingdom
alex.mccaffrey@city.ac.uk

Eduardo Alonso
City, University of London
London, United Kingdom
eduardo.alonso@city.ac.uk

Esther Mondragon
City, University of London
London, United Kingdom
esther.mondragon.@city.ac.uk

ABSTRACT

Efficient exploration remains a challenge in reinforcement learning (RL), especially in stochastic or complex environments. We introduce Predictive Improvement through Latent space OpTimisation (PILOT), an intrinsically motivated RL algorithm that rewards actions leading to improvements in the agent's environmental dynamics model. PILOT optimizes an intrinsic reward signal based on epistemic uncertainty reduction, thereby encouraging structured exploration. Our evaluations against benchmark intrinsic motivation algorithms in challenging environments show that PILOT achieves superior performance and exhibits robustness to stochastic distractions.

ACM Reference Format:
Alexander McCaffrey, Eduardo Alonso, and Esther Mondragon. 2025. Predictive Improvement through Latent Space Optimisation: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Exploration in reinforcement learning can be driven by injecting noise into the action space [5, 7], maximizing state/action entropy [4, 16], setting intermediate goals, or using *intrinsic motivation* [1, 2, 8]. Intrinsically motivated RL augments environmental rewards with intrinsic signals to encourage state space exploration, often leveraging prediction errors from neural networks to quantify novelty [3]. However, using raw prediction errors alone has limitations [2, 9]. High errors do not always indicate learnable dynamics, especially in stochastic environments where errors may remain persistently high despite repeated sampling. This can mislead the agent into exploring inherently unpredictable regions rather than learning useful transitions. A more effective approach is to reward *prediction improvement* [11, 12]. Instead of rewarding absolute error, this method incentivizes transitions where the agent's model improves after sampling, focusing on reducible epistemic uncertainty rather than stochastic noise. We introduce PILOT, which shapes intrinsic rewards based on prediction improvements in a latent space learned via an inverse dynamics model.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 BACKGROUND

Exploration in reinforcement learning is essential for discovering optimal policies, yet it remains a significant challenge. Traditional exploration strategies, such as epsilon-greedy and entropy-based approaches, often fail in environments with sparse rewards or high stochasticity. To address this, intrinsic motivation has been widely adopted as a mechanism to encourage structured exploration.Intrinsic motivation techniques enhance the agent's reward function by incorporating internal signals based on novelty, curiosity, or prediction errors. Prediction error-based intrinsic rewards, in particular, have gained traction, leveraging the observation that neural networks struggle to generalize to unseen states, resulting in higher prediction errors. By rewarding the agent for encountering states with high prediction error, these methods encourage exploration beyond immediate extrinsic rewards.

However, using prediction error alone as an intrinsic reward has notable pitfalls. High prediction error does not always indicate learnable dynamics, especially in stochastic environments where errors remain persistently high due to randomness rather than meaningful uncertainty. This can lead to inefficient exploration, as agents may be incentivized to visit highly unpredictable regions rather than focusing on acquiring useful knowledge about the environment. To overcome these issues, prediction improvement has emerged as a more effective alternative. Instead of rewarding raw prediction error, this approach incentivizes reductions in prediction error after the model updates, ensuring the agent focuses on epistemic uncertainty—uncertainty that can be resolved through further interaction. This framework underpins PILOT, allowing it to guide exploration towards meaningful state transitions while avoiding distractions from irreducible stochasticity.

3 METHOD

PILOT generates intrinsic rewards by leveraging improvements in prediction quality within a structured feature space. Using raw sensory inputs for intrinsic rewards is suboptimal due to high-dimensional noise and irrelevant factors [2, 9]. Instead, PILOT employs an inverse dynamics model to learn a feature representation that retains action-dependent features while filtering out stochasticity.

The inverse dynamics model predicts actions given consecutive states:

$$\hat{a}_t = g(s_t, s_{t+1}; \theta_I) \tag{1}$$

where θ_I is trained to minimize:

$$\min_{\theta_I} L_I(\hat{a}_t, a_t) \tag{2}$$

ensuring the learned feature space, $\phi(s_t)$, captures controllable environment aspects [14]. The forward model further refines learning

by predicting next encoded states:

$$\hat{\phi}(s_{t+1}) = f(\phi(s_t), a_t; \theta_F) \tag{3}$$

where θ_F are the parameters of the forward model.

Intrinsic rewards are based on reducing prediction errors, distinguishing between:

(1) Learnt Transitions: Transitions that exhibit dynamics the agent has already learnt. These generate very low prediction error and do not represent an opportunity for the agent to improve.
(2) Learnable Transitions: Transitions that exhibit dynamics that have not yet been learnt by the agent and hence generate high prediction error. However, these transitions can be said to be on the border of the agent's knowledge, and this error can be reduced through further sampling.
(3) Unlearnable Transitions: Transitions that exhibit dynamics which, for whatever reason (e.g., too complex or highly stochastic), are not currently learnt by the agent and cannot be learnt quickly.

We note that as both **2** and **3** exhibit dynamics not yet learned by the agent, they can both be expected to generate high prediction error. However, by rewarding *improvement* in this error term we induce a preference for sampling **2** as opposed to **3**.

4 EXPERIMENTAL SETUP

We evaluate PILOT in Gymnasium and DeepMind Control Suite, two standard RL benchmarks. Gymnasium offers diverse tasks, while DeepMind Control Suite focuses on physics-based continuous control. However, both lack real-world perceptual challenges, such as visual distractions. To address this, we introduce noisy environments to assess PILOT's robustness against task-irrelevant stochasticity, comparing its performance with other intrinsic motivation methods.

Evaluating intrinsically motivated agents requires distinguishing between epistemic uncertainty (knowledge gaps) and aleatoric uncertainty (uncontrollable randomness). Effective agents should focus on learnable dynamics while ignoring stochastic noise. To test this, we modify standard RL environments by adding five purely stochastic noise dimensions to the state space, which do not affect dynamics or rewards. This setup assesses whether PILOT prioritizes meaningful exploration over task-irrelevant stochasticity, a key factor for real-world applicability.

Each agent is trained for 1 million timesteps across five independent trials with different random seeds to ensure robustness. Performance is evaluated every 50,000 timesteps over 10 episodes, averaging results to track learning progress. Final results are aggregated across seeds, with the best evaluation score serving as the primary performance metric. We use the Proximal Policy Optimization (PPO) [13] implementation from Stable-Baselines3 [10].

5 RESULTS

The scores of the best-performing policy for each agent (aggregated over 10 environment episodes) are presented in Table 1. Our results demonstrate significant improvements of PILOT over several benchmark intrinsically motivated agents.

Environment	PILOT	ICM	RE3	RND	L2	Baseline
BipedalWalker-v3	**311.019**	304.539	303.405	301.351	61.687	310.062
BipedalWalkerHardcore-v3	**-11.067**	-18.134	-24.224	-27.473	-35.119	-11.071
LunarLanderContinuous-v2	273.54	272.065	263.274	267.89	261.983	**282.58**
MountainCarContinuous-v0	97.171	0.0	**98.768**	97.899	97.486	0.0
Pendulum-v1	-69.406	**-48.719**	-71.283	-48.906	-46.676	-73.521
hopper_hop	**58.219**	25.907	4.288	9.224	7.456	2.913
hopper_stand	**397.055**	144.715	105.653	86.328	124.08	363.261
manipulator_bring_ball	**17.904**	9.805	13.907	5.567	14.912	3.838
point_mass_hard	**596.154**	345.088	56.984	345.943	544.251	150.051
reacher_easy	957.8	960.0	**987.4**	962.4	506.2	968.2
walker_stand	398.037	535.443	**572.018**	562.938	411.578	397.475
walker_walk	403.359	390.891	280.981	**440.523**	245.207	342.308

Table 1: Performance comparison across standard environments

Environment (with Noise)	PILOT	ICM	RE3	RND	L2	Baseline
LunarLanderContinuous-v2	**270.396**	211.982	246.971	265.003	210.121	262.845
MountainCarContinuous-v0	-0.041	-0.109	-1.77	-0.082	-1.708	**-0.011**
Pendulum-v1	**-52.969**	-665.965	-115.139	-545.236	-103.708	-338.459
reacher_easy	399.8	**509.8**	441.8	458.8	278.4	491.0
walker_stand	380.562	430.163	**530.45**	442.514	395.793	390.367
walker_walk	**294.153**	249.099	243.292	291.248	251.867	240.683

Table 2: Performance comparison across distractor environments

In standard environments, PILOT outperforms all other agents in 6 out of 12 tasks, with the next best agent, RE3, leading in only 3. Notably, PILOT is the only agent matching or exceeding baseline performance in complex environments like BipedalWalker-v3 and BipedalWalkerHardcore-v3, supporting our hypothesis that reducing epistemic uncertainty mitigates distractions from intrinsic rewards. To test robustness, we introduced stochastic noise dimensions as distractions for environments in which PILOT was not the best performing agent. Table 2 shows PILOT achieved the best performance in 3 of the 6 environments after noise was added. Unlike ICM and RE3, which suffered substantial performance drops, PILOT maintained or improved its standing, as seen in Pendulum-v1, LunarLanderContinuous-v2, and walker walk. These results highlight PILOT's resilience to task-irrelevant noise, demonstrating its ability to focus on meaningful state-space components.

6 CONCLUSION

Our experimental results demonstrate that PILOT outperforms several competitive baselines, particularly in stochastic settings, by effectively ignoring irrelevant noise dimensions and focusing on meaningful state representations. Notably, PILOT showed significant improvement over other agents when tested on environments with added stochastic dimensions, further highlighting its ability to focus on reducing actionable uncertainty. Future research could expand upon the findings of this work by applying PILOT to more complex and visually rich datasets, such as the Kinetics dataset [6] or the Continuous Distraction Suite [15]. These datasets provide challenging benchmarks for assessing robustness to noise and distractions, which are crucial for scaling reinforcement learning methods to real-world scenarios.

REFERENCES

[1] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. 2016. Unifying Count-Based Exploration and Intrinsic Motivation. *CoRR* abs/1606.01868 (2016), 1479 – 1487.

[2] Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. 2019. Large-Scale Study of Curiosity-Driven Learning. In *7th International Conference on Learning Representations (ICLR 2019)*. ICLR, New Orleans, USA, 1–17. https://iclr.cc/ Seventh International Conference on Learning Representations, ICLR 2019 ; Conference date: 06-05-2019 Through 09-05-2019.

[3] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by random network distillation. In *7th International Conference on Learning Representations (ICLR 2019)*. ICLR, New Orleans, USA, 1–17. https://iclr.cc/ Seventh International Conference on Learning Representations, ICLR 2019 ; Conference date: 06-05-2019 Through 09-05-2019.

[4] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 1352–1361.

[5] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. , 1861–1870 pages.

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV] https://arxiv.org/abs/1705.06950

[7] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1509.02971

[8] Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V. Hafner. 2007. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation* 11, 2 (2007), 265–286. https://doi.org/10.1109/TEVC.2006.890271

[9] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 16–17.

[10] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. http://jmlr.org/papers/v22/20-1364.html

[11] Jürgen Schmidhuber. 1991. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the international conference on simulation of adaptive behavior: From animals to animats*. 222–227.

[12] Jürgen Schmidhuber. 2006. Developmental Robotics, Optimal Artificial Curiosity, Creativity, Music, and the Fine Arts. In *Connection Science*, Vol. 18. 173–187. https://doi.org/10.1080/09540090600768658

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). http://dblp.uni-trier.de/db/journals/corr/corr1707.html#SchulmanWDRK17

[14] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. 2015. Incentivizing exploration in reinforcement learning with deep predictive models. *Advances in Neural Information Processing Systems* (2015), 1353–1361.

[15] Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. 2021. The Distracting Control Suite – A Challenging Benchmark for Reinforcement Learning from Pixels. https://doi.org/10.48550/arXiv.2101.02722

[16] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. 2008. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1433–1438.