

Coherence-Seeking Reinforcement Learning with Intrinsic Creativity

Armando Vieira, University Tartu

Jan 2026

Abstract

Intrinsic motivation methods in reinforcement learning (RL) typically reward novelty or prediction error. These signals can improve exploration but often conflate meaningful discovery with irreducible stochasticity and may incentivize chaotic behavior. We propose an alternative formulation of intrinsic creativity as *coherence formation*: an agent seeks states where its internally generated expectations are challenged, and is rewarded for resolving this mismatch into stable predictive structure.

We introduce a reach–yield decomposition. *Reach* is an intention-conditioned predictive model of what the agent is prepared to bring forth through action; *yield* is a separately learned empirical model of what the environment returns under the agent’s attempted interactions. We define *incoherence* as the divergence between reach and yield, and construct intrinsic rewards that (i) regulate interaction near a target mismatch band (“creative edge”) and (ii) reward reduction of incoherence over time (coherence formation). We present a practical actor–critic training algorithm with stable variants (distributional KL and an MSE surrogate), provide diagnostics on CartPole, and position CSRL relative to major intrinsic motivation families.

1 Introduction

Exploration and representation learning remain core challenges in reinforcement learning, especially when extrinsic reward is sparse or delayed. A common approach adds an intrinsic reward signal that encourages novelty, surprise, or prediction error. Examples include prediction-error curiosity [1], random network distillation [2], pseudo-count bonuses [3], episodic novelty [5], and directed exploration curricula [6].

However, novelty and prediction error are imperfect proxies for discovery. Prediction errors can remain high in stochastic or chaotic regions, and novelty can overemphasize irrelevant variation. Human creativity suggests a different pattern: temporary mismatch (tension) followed by reorganization into a new stable understanding. We operationalize this as *intrinsic creativity*: seek *productive* mismatch and reward its *resolution*.

This paper introduces a reach–yield decomposition that makes this notion explicit and algorithmic. The key idea is not to maximize error, but to regulate mismatch around a target range and reward the process of becoming coherent.

2 Preliminaries

We consider an MDP $(\mathcal{S}, \mathcal{A}, p, r^{\text{ext}}, \gamma)$ with states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, transition kernel $p(s' | s, a)$, extrinsic reward $r^{\text{ext}}(s, a, s')$, and discount $\gamma \in (0, 1)$.

We optionally introduce latent intentions/skills $z \in \mathcal{Z}$ sampled from $p(z | s)$ or fixed per episode. The control policy is $\pi_\theta(a | s, z)$.

We define an *outcome representation* $o = g_\eta(s') \in \mathbb{R}^d$ (e.g., a learned embedding of the next state, a goal vector, or task-relevant coordinates). Reach and yield are compared in this outcome space.

3 Reach–Yield Decomposition

3.1 Reach

Reach represents what the agent expects it can bring forth through action under intention z . We model reach as a conditional outcome distribution

$$q_\theta(o \mid s, z). \tag{1}$$

A common instantiation is diagonal Gaussian:

$$q_\theta(o \mid s, z) = \mathcal{N}(\mu_r(s, z), \text{diag}(\sigma_r^2(s, z))). \tag{2}$$

Intuitively, reach is “agent-shaped”: it reflects competence, inductive biases, and intention conditioning.

3.2 Yield

Yield represents the environment’s empirical constraints on outcomes when the agent attempts action a under intention z . We model yield as

$$y_\phi(o \mid s, z, a), \tag{3}$$

trained to match observed outcomes. A diagonal Gaussian instantiation is

$$y_\phi(o \mid s, z, a) = \mathcal{N}(\mu_y(s, z, a), \text{diag}(\sigma_y^2(s, z, a))). \tag{4}$$

Yield is “world-shaped”: it is trained primarily by supervised prediction from experience.

Architectural separation. Reach and yield should not be forced to agree (e.g., via shared heads). Their mismatch is treated as a signal.

4 Incoherence: Reach–Yield Divergence

We define *incoherence* as a divergence between reach and yield.

4.1 Distributional incoherence (KL)

For distributional reach and yield, we define

$$I_{\text{KL}}(s, z, a) = D_{\text{KL}}(q_\theta(o \mid s, z), \mid, y_\phi(o \mid s, z, a)). \tag{5}$$

If both are diagonal Gaussians, KL has closed form:

$$D_{\text{KL}}(p \mid q) = \frac{1}{2} \sum_{i=1}^d \left[\log \frac{\sigma_{q,i}^2}{\sigma_{p,i}^2} + \frac{\sigma_{p,i}^2 + (\mu_{p,i} - \mu_{q,i})^2}{\sigma_{q,i}^2} - 1 \right]. \tag{6}$$

4.2 Stable surrogate incoherence (MSE)

Flexible Gaussian yield models can collapse variance and yield unbounded KL. A stable surrogate compares predictive means:

$$I_{\text{MSE}}(s, z, a) = \|\mu_r(s, z) - \mu_y(s, z, a)\|_2^2. \quad (7)$$

This preserves the reach–yield mismatch interpretation while avoiding density-model pathologies.

5 Intrinsic Creativity Reward

Intrinsic reward is constructed from incoherence in two parts.

5.1 Creative-edge regulation

We encourage operating near a target mismatch I^* :

$$r_{\text{edge}}(t) = \exp\left(-\frac{(I_t - I^*)^2}{2\sigma_I^2}\right), \quad (8)$$

where $I_t = I(s_t, z_t, a_t)$.

5.2 Coherence formation

We reward reductions in incoherence:

$$r_{\text{form}}(t) = \text{clip}(\max(0, I_t - I_{t+1}), 0, c). \quad (9)$$

The clip constant c prevents rare large drops from dominating.

5.3 Total intrinsic reward

$$r^{\text{int}}(t) = \alpha r_{\text{edge}}(t) + \beta r_{\text{form}}(t), \quad (10)$$

with small α, β . Total reward is

$$r(t) = r^{\text{ext}}(t) + r^{\text{int}}(t). \quad (11)$$

6 Training: Actor–Critic with Advantage Shaping

We describe a practical on-policy actor–critic procedure.

6.1 Yield learning

Given transitions (s_t, z_t, a_t, s_{t+1}) , compute $o_t = g_\eta(s_{t+1})$.

Gaussian yield (MLE).

$$\min_{\phi} \mathbb{E} [-\log y_\phi(o_t | s_t, z_t, a_t)]. \quad (12)$$

MSE yield (stable baseline).

$$\min_{\phi} \mathbb{E} [\|\mu_y(s_t, z_t, a_t) - o_t\|_2^2]. \quad (13)$$

Algorithm 1 Coherence-Seeking Reinforcement Learning (CSRL)

- 1: Initialize (θ, ψ) for policy/value, ϕ for yield, and outcome map g_η
 - 2: **for** iterations $k = 1, 2, \dots$ **do**
 - 3: Collect on-policy rollouts: $a_t \sim \pi_\theta(\cdot \mid s_t, z_t)$
 - 4: Compute outcomes $o_t \leftarrow g_\eta(s_{t+1})$
 - 5: Update yield ϕ by minimizing MSE or NLL on (s_t, z_t, a_t, o_t)
 - 6: Compute incoherence I_t (KL or MSE) and intrinsic reward r_t^{int}
 - 7: Compute \hat{A}_t^{ext} (GAE on r^{ext}) and \hat{A}_t^{int} (GAE on r^{int})
 - 8: Set shaped advantage $\hat{A} \leftarrow \hat{A}^{\text{ext}} + \lambda \hat{A}^{\text{int}}$
 - 9: Update policy θ with entropy bonus using \hat{A}
 - 10: Update value ψ to regress extrinsic returns (or total returns)
 - 11: **end for**
-

6.2 Policy/value objectives

Let $V_\psi(s, z)$ be the value function. We compute generalized advantage estimates (GAE) for extrinsic and intrinsic rewards separately.

Let \hat{A}_t^{ext} be GAE using r^{ext} and \hat{A}_t^{int} be GAE using r^{int} . We form a shaped advantage

$$\hat{A}_t = \hat{A}_t^{\text{ext}} + \lambda \hat{A}_t^{\text{int}}, \quad \lambda \in [0, 1]. \quad (14)$$

This keeps task learning primary while letting intrinsic reward bias exploration.

The policy objective (A2C style) is

$$\max_{\theta}; \mathbb{E} \left[\log \pi_\theta(a_t \mid s_t, z_t), \hat{A} * t \right] + \kappa, \mathcal{H}(\pi * \theta), \quad (15)$$

where \mathcal{H} is entropy.

The value objective (often using extrinsic returns for stability) is

$$\min_{\psi}; \mathbb{E} \left[(V_\psi(s_t, z_t) - \hat{R}_t^{\text{ext}})^2 \right]. \quad (16)$$

6.3 Algorithm

7 Related Work and Comparison

We compare CSRL to major intrinsic motivation families.

7.1 Prediction-error curiosity

ICM [1] rewards forward prediction error in a learned feature space; RND [2] rewards prediction error of a random feature network. These methods are effective but can incentivize irreducible noise. CSRL differs by (i) introducing an explicit reach–yield split and (ii) rewarding *resolution* of mismatch via ΔI rather than rewarding mismatch itself.

7.2 Count- and density-based exploration

Pseudo-count methods generalize visitation counts via density models and turn them into bonuses [3]. CSRL does not target visitation frequency; the signal depends on how intentions and actions map to outcomes under environmental constraints.

7.3 Information gain and Bayesian surprise

VIME [4] rewards information gain about dynamics parameters (a KL between posterior and prior in parameter space). CSRL can be seen as an interaction-level analog: divergence between an intention-conditioned “readiness” distribution and an empirical constraint distribution.

7.4 Episodic novelty and directed exploration

Episodic curiosity uses reachability-based novelty via episodic memory [5]; NGU combines episodic and lifelong novelty with learned exploration policies [6]. CSRL is complementary and could incorporate memory by making yield history-dependent (e.g., via recurrent or reservoir predictors).

7.5 Skill discovery and mutual information

DIAYN [7] and VIC [8] maximize mutual information between skill identity and outcomes to learn diverse behaviors. CSRL is compatible with skill conditioning but focuses on the *formation of coherence*: skills are valuable insofar as they reliably reduce reach–yield mismatch across contexts.

7.6 Active inference and epistemic value

Active inference casts behavior as minimizing expected free energy, balancing epistemic and pragmatic terms [9, 10]. CSRL shares a mismatch intuition but remains in standard RL: intrinsic reward is shaping, not a replacement for reward maximization.

7.7 Learning progress and compression progress

Psychological and theoretical accounts connect curiosity to learning progress [11] and compression progress [12]. CSRL operationalizes a related idea via decreases in reach–yield divergence.

8 Experiments: CartPole Diagnostics (Initial Validation)

We provide an initial validation on CartPole-v1 to test stability and interpretability.

Outcome representation. We used task-relevant outcomes $o_t = (\theta_{t+1}, \dot{\theta}_{t+1})$, biasing coherence toward balance dynamics.

Metrics. We report episode return (not per-step reward). We also report yield prediction error and mean incoherence to diagnose stability.

Qualitative behavior. Yield MSE decreases rapidly; incoherence remains bounded; intrinsic reward stays small. CSRL improves above random performance but may not exceed tuned PPO/A2C on this dense-reward task. The primary contribution is the formulation and stable training recipe.

9 Limitations and Future Work

Outcome choice. The intrinsic signal depends on the outcome representation; poor choices can decouple coherence from task utility.

Distributional calibration. KL-based incoherence is attractive but needs regularization (e.g., fixed variance, priors, or calibrated density estimators) to prevent variance collapse.

Where we expect gains. Sparse rewards, multi-modal skills, and long-horizon credit assignment are the most promising regimes.

10 Conclusion

We proposed CSRL, framing intrinsic creativity as regulated reach–yield mismatch and rewarding its resolution into new coherence. By separating reach and yield and using divergence dynamics as shaping signals, CSRL provides a principled alternative to novelty-based exploration.

References

- [1] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. *ICML*, 2017.
- [2] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *ICLR*, 2019. arXiv:1810.12894.
- [3] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. *NeurIPS*, 2016. arXiv:1606.01868.
- [4] R. Houthoofd, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. VIME: Variational information maximizing exploration. *NeurIPS*, 2016. arXiv:1605.09674.
- [5] N. Savinov, A. Raichuk, R. Marinier, D. Vincent, M. Pollefeys, T. Lillicrap, and S. Gelly. Episodic curiosity through reachability. *ICLR*, 2019. arXiv:1810.02274.
- [6] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, D. Guo, and C. Blundell. Never give up: Learning directed exploration strategies. *ICLR*, 2020. arXiv:2002.06038.
- [7] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *ICLR*, 2019. arXiv:1802.06070.
- [8] K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control. arXiv:1611.07507, 2016.
- [9] N. Sajid, K. Friston, and T. Parr. Active inference: Demystified and compared. *Neural Computation*, 2021 (preprint: [sajid.pdf](https://activeinference.github.io/papers/sajid.pdf)(<https://activeinference.github.io/papers/sajid.pdf>)).
- [10] T. Parr and K. Friston. Generalised free energy and active inference. *Biological Cybernetics*, 2019.
- [11] P.-Y. Oudeyer, J. Gottlieb, and M. Lopes. Intrinsic motivation, curiosity, and learning. *Progress in Brain Research*, 2016.
- [12] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2010.