DeepMind

5

Curiosity in Hindsight

Exploration in Stochastic Environments

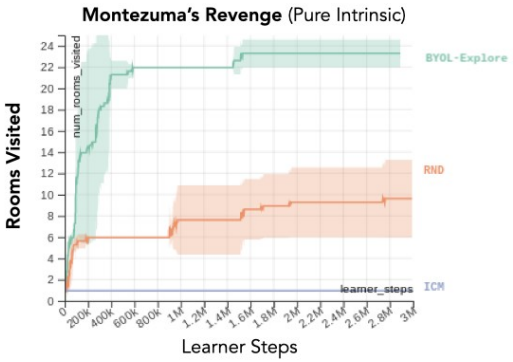
mh/39030, mh/38729, mh/p17043

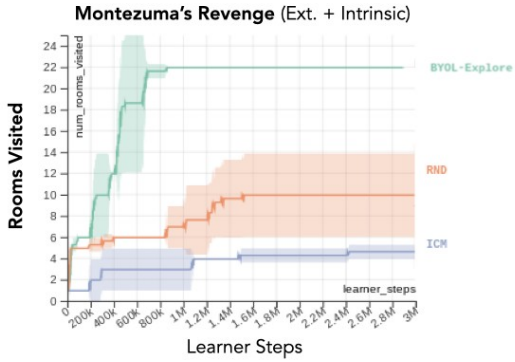
Lead(s): Daniel Jarrett

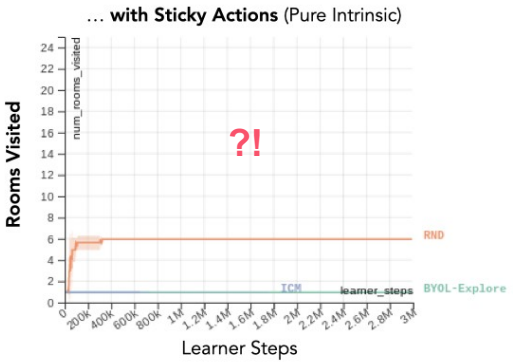
Contributor(s): Corentin Tallec, Florent Altché,
Thomas Mesnard, Kat McKinney, Rémi Munos, Michal Valko

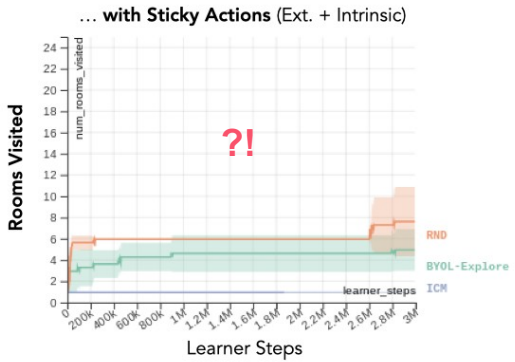
Montezuma's Revenge — Rooms Visited

Stochasticity can break performance of predictive error-based exploration, e.g. BYOL-Explore.



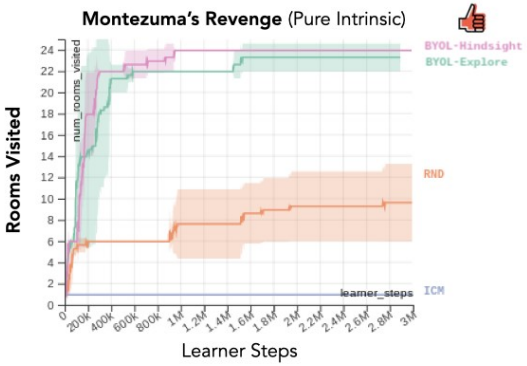


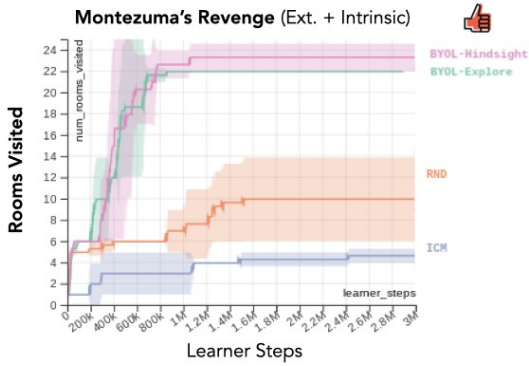


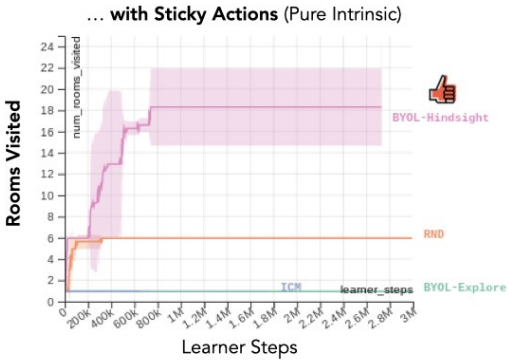


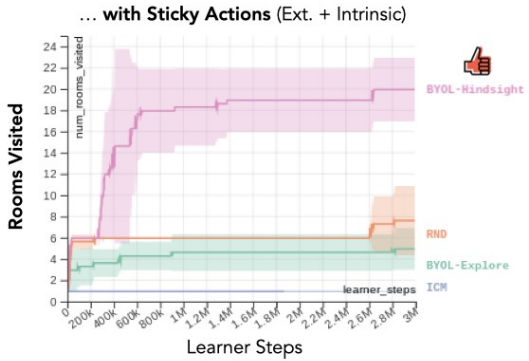
Montezuma's Revenge — Rooms Visited

Hindsight can improve performance of BYOL-Explore, esp. in high-stochasticity settings.



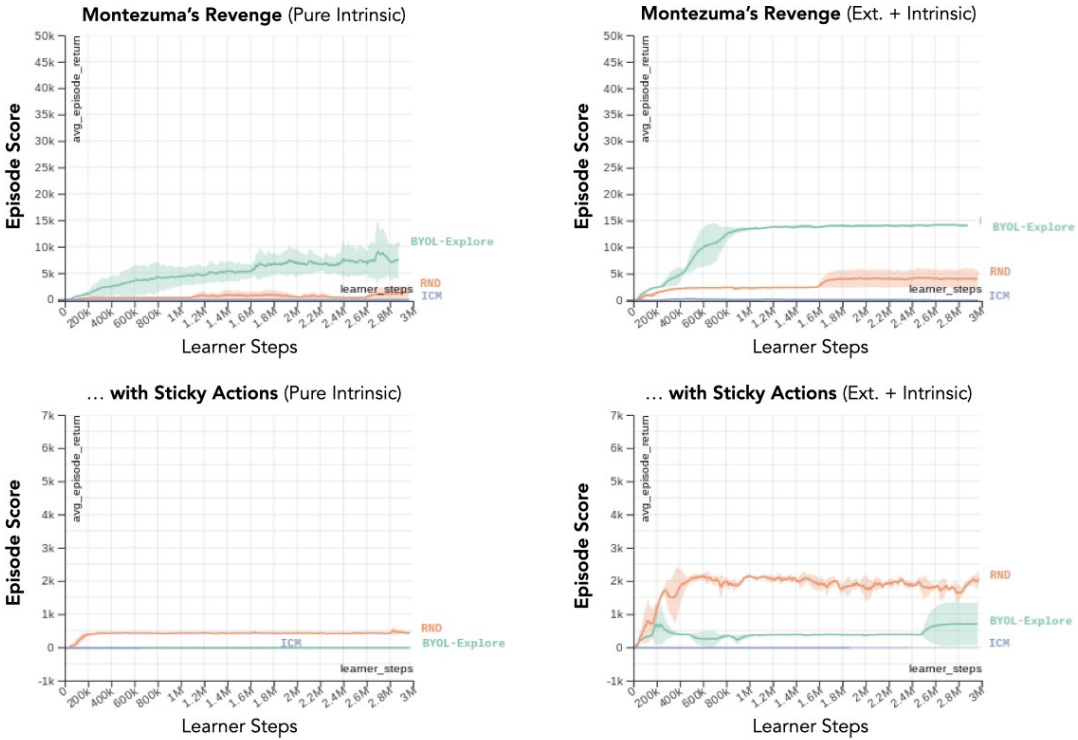






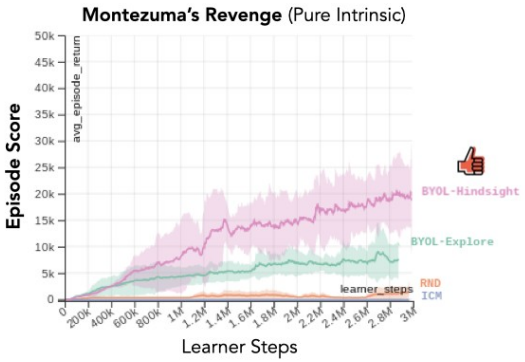
Montezuma's Revenge — Episode Score

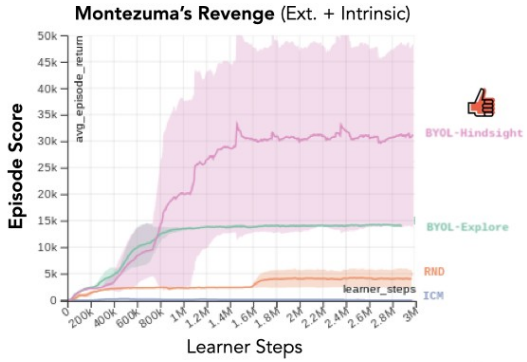
Hindsight can improve performance of BYOL-Explore, esp. in high-stochasticity settings.

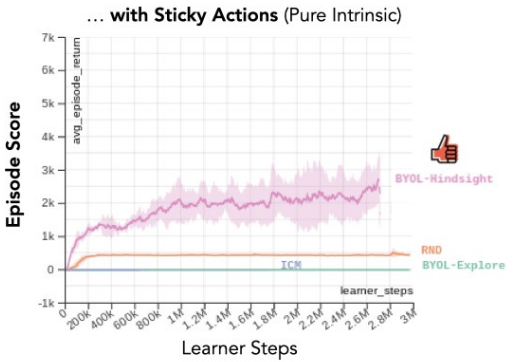


Montezuma's Revenge — Episode Score

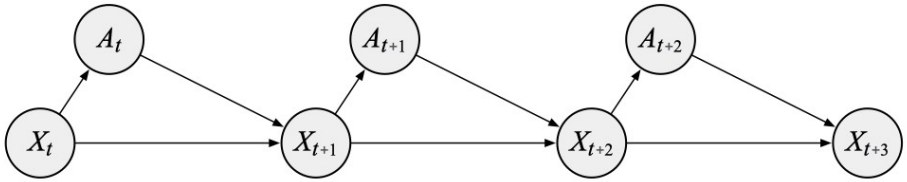
Hindsight can improve performance of BYOL-Explore, esp. in high-stochasticity settings.





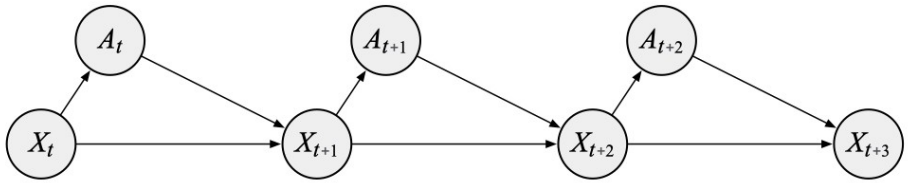


Introduction



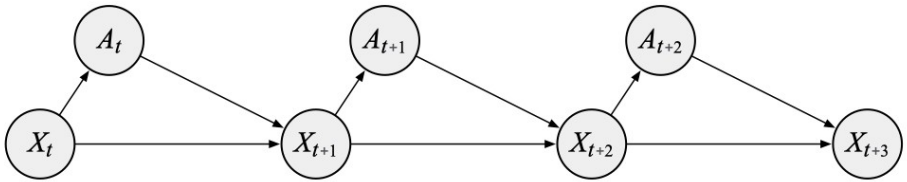
Introduction

Problem: How to explore the world when external rewards are sparse or absent?



Introduction

Problem: How to explore the world when external rewards are sparse or absent?
Curiosity: Prioritize exploring—and learning from—what is not yet understood.

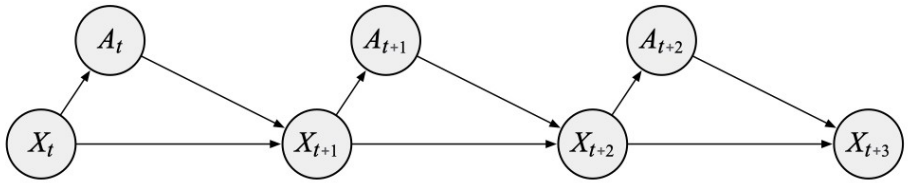


Introduction

Problem: How to explore the world when external rewards are sparse or absent?
Curiosity: Prioritize exploring—and learning from—what is not yet understood.
Predictive Error-based Exploration:

"understanding" := ability to predict outcomes

Dynamics Model → *Outcome Prediction* ∝ *Intrinsic Reward* → *Agent Policy*



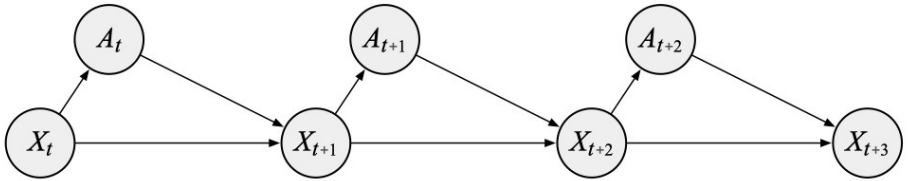
Introduction

Problem: How to explore the world when external rewards are sparse or absent?
Curiosity: Prioritize exploring—and learning from—what is not yet understood.
Predictive Error-based Exploration:

"understanding" := ability to predict outcomes

Dynamics Model → *Outcome Prediction* ∝ *Intrinsic Reward* → *Agent Policy*

Hurdle 1: **Dimensionality**.



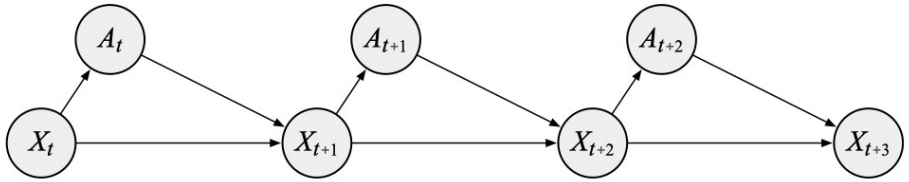
Introduction

Problem: How to explore the world when external rewards are sparse or absent?
Curiosity: Prioritize exploring—and learning from—what is not yet understood.
Predictive Error-based Exploration:

"understanding" := ability to predict outcomes

Dynamics Model → *Outcome Prediction* ∝ *Intrinsic Reward* → *Agent Policy*

Hurdle 1: **Dimensionality**. Solution: **Context Representations**.



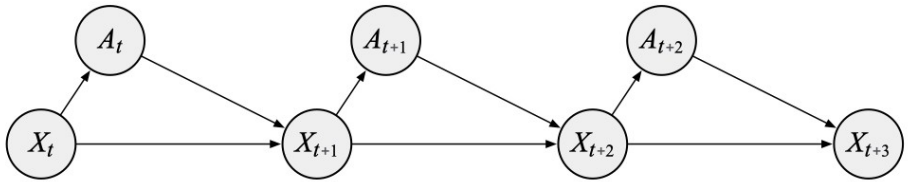
Introduction

Problem: How to explore the world when external rewards are sparse or absent?
Curiosity: Prioritize exploring—and learning from—what is not yet understood.
Predictive Error-based Exploration:

"understanding" := ability to predict outcomes

Dynamics Model → *Outcome Prediction* ∝ *Intrinsic Reward* → *Agent Policy*

Hurdle 1: **Dimensionality**. Solution: **Context Representations**.
Hurdle 2: **Stochasticity**.



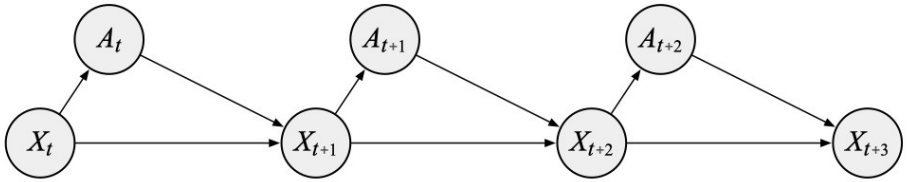
Introduction

Problem: How to explore the world when external rewards are sparse or absent?
Curiosity: Prioritize exploring—and learning from—what is not yet understood.
Predictive Error-based Exploration:

"understanding" := ability to predict outcomes

Dynamics Model → *Outcome Prediction* ∝ *Intrinsic Reward* → *Agent Policy*

Hurdle 1: **Dimensionality**. Solution: **Context Representations**.
Hurdle 2: **Stochasticity**. *epistemic knowledge* (viz. necessary truths) ← "novelty"
vs. *aleatoric variation* (viz. contingent facts) ← "noise"



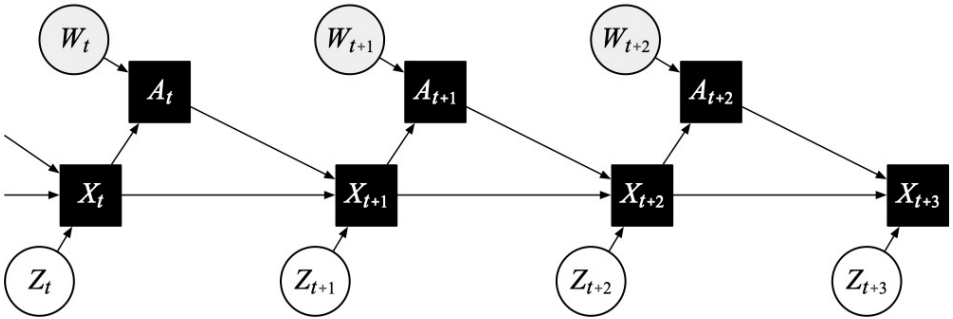
Introduction

Problem: How to explore the world when external rewards are sparse or absent?
Curiosity: Prioritize exploring—and learning from—what is not yet understood.
Predictive Error-based Exploration:

"understanding" := ability to predict outcomes

Dynamics Model → *Outcome Prediction* ∝ *Intrinsic Reward* → *Agent Policy*

Hurdle 1: **Dimensionality**. Solution: **Context Representations**.
Hurdle 2: **Stochasticity**. Solution: **Hindsight Representations**.



Overview

1. **Introduction**: Curiosity-driven Exploration
2. **Motivation**: Stochastic Environments
3. **Curiosity in Hindsight**
4. **Practical Framework**
5. **Experiments**
6. **Future Directions**

Overview

1. **Introduction**: Curiosity-driven Exploration
2. **Motivation**: Stochastic Environments
3. **Curiosity in Hindsight**
4. **Practical Framework**
5. **Experiments**
6. **Future Directions**

Motivation — Problem Formalism

Notation: states X, actions A, dynamics τ, policy π.

Motivation — Problem Formalism

Notation: states X, actions A, dynamics τ, policy π.

Curiosity: Define the intrinsic reward

$$\underset{pred.}{\text{Reward}}\ (x_t, a_t) := \mathbb{E}_{X_{t+1}} \left\| X_{t+1} - \hat{x}_{t+1} \right\|_2^2$$

The agent performs

$$\underset{\pi}{\text{(policy) maximize}}\ \ \underset{\hat{\tau}}{\text{(model) min}}\ \ \mathbb{E}_{X_t, A_t}\left[\underset{pred.}{\text{Reward}}\ (X_t, A_t) \right]$$

Motivation — Problem Formalism

Notation: states X, actions A, dynamics τ, policy π.

Curiosity: Define the intrinsic reward

$$\underset{\textit{pred.}}{\text{Reward}}\ (x_t, a_t) := \mathbb{E}_{X_{t+1}} \left\| X_{t+1} - \hat{x}_{t+1} \right\|_2^2$$

The agent performs

$$\underset{\pi}{\text{(policy) maximize}}\ \ \underset{\hat{\tau}}{\text{(model) min}}\ \ \mathbb{E}_{X_t, A_t}\left[\underset{\textit{pred.}}{\text{Reward}}\ (X_t, A_t) \right]$$

Stochastic Traps: The reward converges to the entropy!

Motivation — Problem Formalism

Notation: states X, actions A, dynamics τ, policy π.

Curiosity: Define the intrinsic reward

$$\underset{\textit{pred.}}{\text{Reward}}\ (x_t, a_t) := \mathbb{E}_{X_{t+1}} \left\| X_{t+1} - \hat{x}_{t+1} \right\|_2^2$$

The agent performs

$$\underset{\pi}{\text{(policy) maximize}}\ \ \underset{\hat{\tau}}{\text{(model) min}}\ \ \mathbb{E}_{X_t, A_t}\left[\underset{\textit{pred.}}{\text{Reward}}\ (X_t, A_t) \right]$$

Stochastic Traps: The reward converges to the entropy!

What about…

$$\text{Reward}\ (x_t, a_t) \overset{?}{:=} \text{Divergence}\Big(\underbrace{\tau(X_{t+1} \mid x_t, a_t)}_{\text{(reality)}} \ \Big\|\ \underbrace{\hat{\tau}(X_{t+1} \mid x_t, a_t)}_{\text{(model)}} \Big)$$

…but how?

Motivation — Related Work

Curiosity-driven Exploration Method	*Prediction Inputs*	*Prediction Target*	*Measure of Learning*	Random Noise	X-/A-Dep. Noise	Dynamics Awareness	**Representation Space**
AE [10]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	reconstructive
ICM [11]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✓	✗	✓	action predictive
EMI [13]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	MI-maximizing
RND [12]	X_t	$f_{\text{random}}(X_t)$	$\mathcal{L}_\eta^{\text{predict}}$	✓	✓	✗	random projection
Dora [16]	X_t, A_t	const. zero	$\mathcal{L}_\eta^{\text{predict}}$	✓	✓	✗	pixel space
AMA [15]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}} - \text{Tr}(\hat{\Sigma}_{t+1})$	✓	✓	✓	pixel space
BYOL-Explore [14]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	bootstrapped
Curiosity in Hindsight + any representation	X_t, A_t, Z_{t+1}	X_{t+1}	$\mathcal{L}_{\theta,\eta}^{\text{reconstruct}} + \mathcal{L}_{\theta,\nu}^{\text{invariance}}$	✓	✓	✓	***any representation***

Existing methods are either tied to specific representations, susceptible to some noise, or a significant departure from the dynamics-aware notion of curiosity.

[10] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *International Conference on Learning Representations*, 2016.

[11] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[12] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *International Conference on Learning Representations*, 2019.

[13] Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. Emi: Exploration with mutual information. In *International Conference on Machine Learning*, pages 3360–3369. PMLR, 2019.

[14] Zhaohan Daniel Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35, 2022.

[15] Augustine Mavor-Parker, Kimberly Young, Caswell Barry, and Lewis Griffin. How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation. In *International Conference on Machine Learning*, pages 15220–15240. PMLR, 2022.

[16] Leshem Choshen, Lior Fox, and Yonatan Loewenstein. Dora the explorer: Directed outreaching reinforcement action-selection. *International Conference on Learning Representations*, 2018.

Motivation — Related Work

Curiosity-driven Exploration Method	*Prediction Inputs*	*Prediction Target*	*Measure of Learning*	Random Noise	X-/A-Dep. Noise	Dynamics Awareness	**Representation Space**
AE [10]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	reconstructive
ICM [11]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✓	✗	✓	action predictive
EMI [13]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	MI-maximizing
RND [12]	X_t	$f_{\text{random}}(X_t)$	$\mathcal{L}_\eta^{\text{predict}}$	✓	✓	✗	random projection
Dora [16]	X_t, A_t	const. zero	$\mathcal{L}_\eta^{\text{predict}}$	✓	✓	✗	pixel space
AMA [15]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}} - \text{Tr}(\hat{\Sigma}_{t+1})$	✓	✓	✓	pixel space
BYOL-Explore [14]	X_t, A_t	X_{t+1}	$\mathcal{L}_\eta^{\text{predict}}$	✗	✗	✓	bootstrapped
Curiosity in Hindsight + any representation	X_t, A_t, Z_{t+1}	X_{t+1}	$\mathcal{L}_{\theta,\eta}^{\text{reconstruct}} + \mathcal{L}_{\theta,\nu}^{\text{invariance}}$	✓	✓	✓	***any representation***

Existing methods are either tied to specific representations, susceptible to some noise, or a significant departure from the dynamics-aware notion of curiosity.

Curiosity in Hindsight

1. **Stochasticity Types**: It handles different types of stochasticities in generality.
2. **Dynamics Awareness**: It does not discard the curiosity-driven paradigm entirely.
3. **Generality and Scalability**: It is not be restrictive wrt. representations/computation.

Overview

1. **Introduction**: Curiosity-driven Exploration
2. **Motivation**: Stochastic Environments
3. **Curiosity in Hindsight**
4. **Practical Framework**
5. **Experiments**
6. **Future Directions**

Example — Betting on a Hidden Coin

We take the action A_t = "bet on heads".

We observe the outcome X_{t+1} = "lost the bet".

Example — Betting on a Hidden Coin

We take the action A_t = "bet on heads".
We observe the outcome X_{t+1} = "lost the bet".

Two facts are clear:
1. *a priori*, we could not have predicted this result at all;
2. *a posteriori*, we may deduce the latent fact Z_{t+1} = "coin landed tails".

Example — Betting on a Hidden Coin

We take the action A_t = "bet on heads".
We observe the outcome X_{t+1} = "lost the bet".

Two facts are clear:
1. *a priori*, we could not have predicted this result at all;
2. *a posteriori*, we may deduce the latent fact Z_{t+1} = "coin landed tails".

Knowing how the game works:
In hindsight (i.e. given Z_{t+1}), the outcome is obvious (i.e. X_{t+1} is identified).

Not knowing how the game works:
We wouldn't be able to infer Z_{t+1}, nor would having it let us identify X_{t+1}.

Example — Betting on a Hidden Coin

We take the action A_t = "bet on heads".
We observe the outcome X_{t+1} = "lost the bet".

Two facts are clear:
1. *a priori*, we could not have predicted this result at all;
2. *a posteriori*, we may deduce the latent fact Z_{t+1} = "coin landed tails".

Knowing how the game works:
In hindsight (i.e. given Z_{t+1}), the outcome is obvious (i.e. X_{t+1} is identified).

Not knowing how the game works:
We wouldn't be able to infer Z_{t+1}, nor would having it let us identify X_{t+1}.

Curiosity:

"understanding" := ability to predict outcomes *a priori*

Example — Betting on a Hidden Coin

We take the action A_t = "bet on heads".
We observe the outcome X_{t+1} = "lost the bet".

Two facts are clear:
1. *a priori*, we could not have predicted this result at all;
2. *a posteriori*, we may deduce the latent fact Z_{t+1} = "coin landed tails".

Knowing how the game works:
In hindsight (i.e. given Z_{t+1}), the outcome is obvious (i.e. X_{t+1} is identified).

Not knowing how the game works:
We wouldn't be able to infer Z_{t+1}, nor would having it let us identify X_{t+1}.

Curiosity:

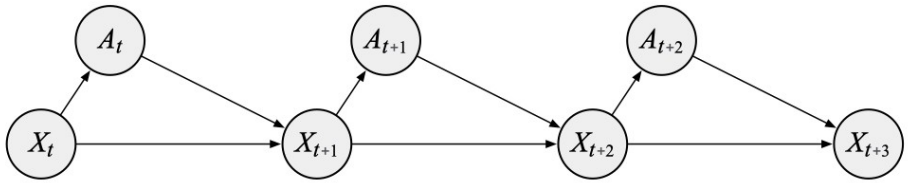
"understanding" := ability to ~~predict~~ outcomes ~~*a priori*~~
reconstruct *a posteriori*

Overview

1. **Introduction**: Curiosity-driven Exploration
2. **Motivation**: Stochastic Environments
3. **Curiosity in Hindsight**
   - Structural Causal Model
   - Hindsight Representations
   - Optimistic Exploration
4. **Practical Framework**
5. **Experiments**
6. **Future Directions**

Curiosity in Hindsight — Structural Causal Graph

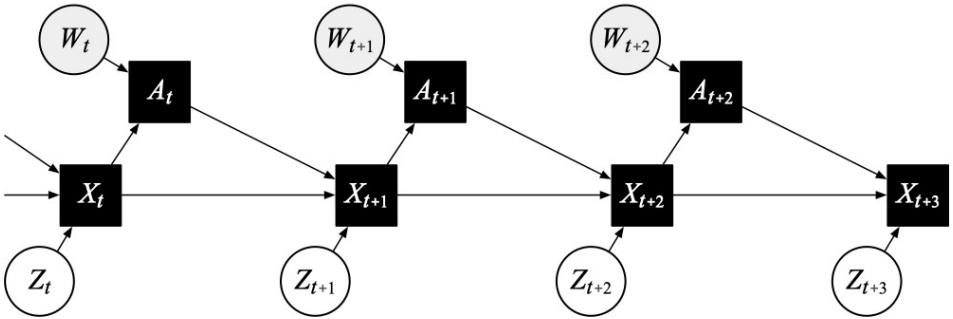
Notation: noise variable Z, encapsulating *all* sources of unobserved stochasticity.



Curiosity in Hindsight — Structural Causal Graph

Notation: noise variable Z, encapsulating *all* sources of unobserved stochasticity.

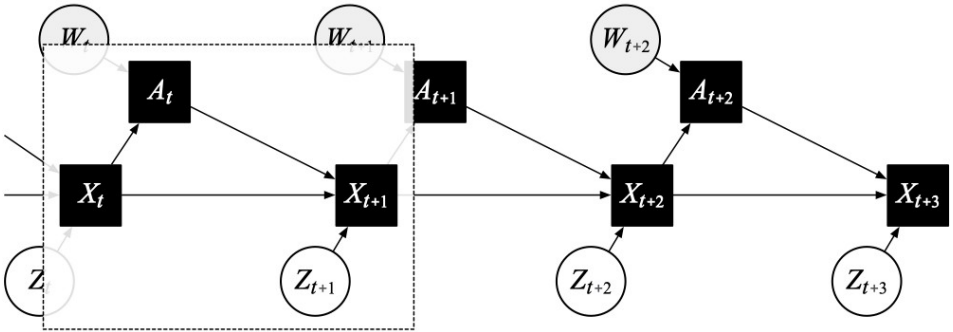
By construction, $x_{t+1} = f(x_t, a_t, z_{t+1})$ for some deterministic function f.



Curiosity in Hindsight — Structural Causal Graph

Notation: noise variable Z, encapsulating *all* sources of unobserved stochasticity.

By construction, $x_{t+1} = f(x_t, a_t, z_{t+1})$ for some deterministic function f.



Curiosity in Hindsight — Structural Causal Graph

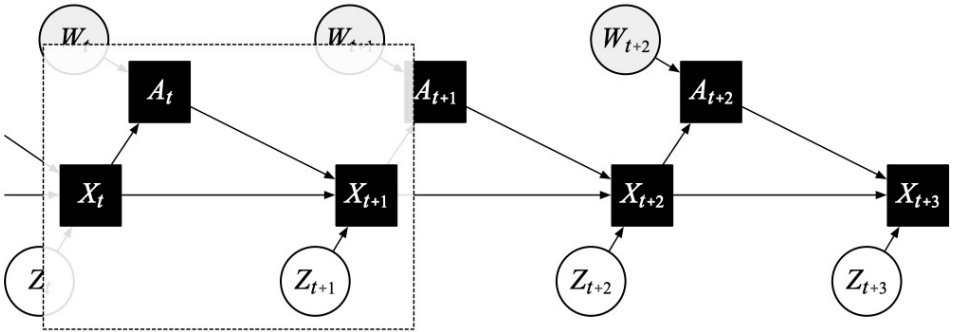
Notation: noise variable Z, encapsulating *all* sources of unobserved stochasticity.

By construction, $x_{t+1} = f(x_t, a_t, z_{t+1})$ for some deterministic function f.

Sample from Posterior

For any $Z_{t+1} \sim p(\cdot \mid x_t, a_t, x_{t+1})$:

$$f(x_t, a_t, Z_{t+1}) = x_{t+1}$$



Curiosity in Hindsight — Structural Causal Graph

Notation: noise variable Z, encapsulating *all* sources of unobserved stochasticity.

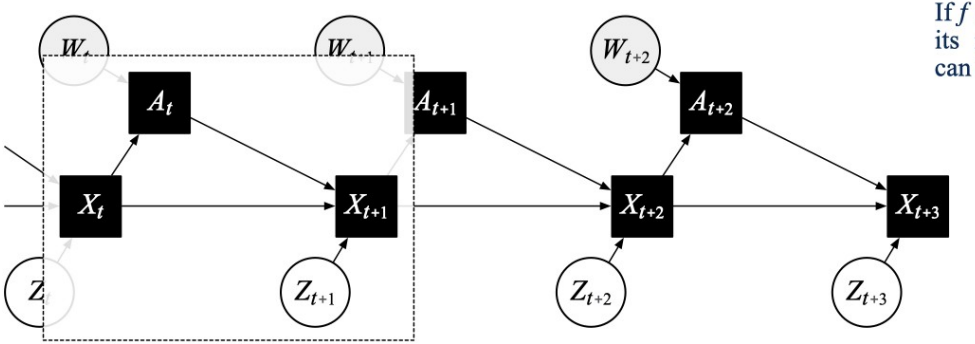
By construction, $x_{t+1} = f(x_t, a_t, z_{t+1})$ for some deterministic function f.

Sample from Posterior

For any $Z_{t+1} \sim p(\cdot \mid x_t, a_t, x_{t+1})$:

$$f(x_t, a_t, Z_{t+1}) = x_{t+1}$$

If f needs to be *learned*, its reconstruction error can be driven to zero…!



Curiosity in Hindsight — Structural Causal Graph

Notation: noise variable Z, encapsulating *all* sources of unobserved stochasticity.

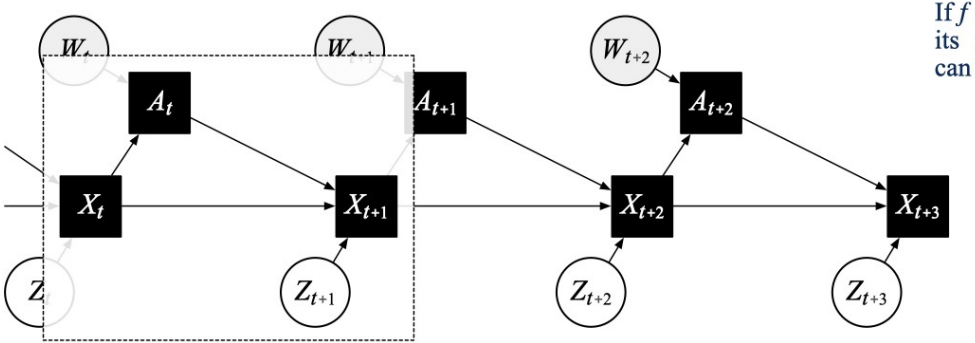
By construction, $x_{t+1} = f(x_t, a_t, z_{t+1})$ for some deterministic function f.

Sample from Posterior

For any $Z_{t+1} \sim p(\cdot \mid x_t, a_t, x_{t+1})$:

$$f(x_t, a_t, Z_{t+1}) = x_{t+1}$$

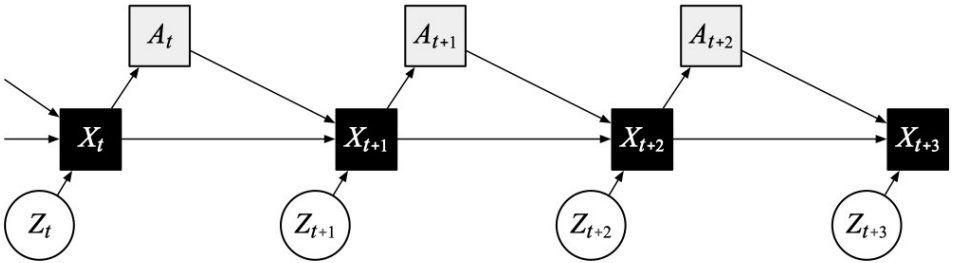
If f needs to be *learned*, its reconstruction error can be driven to zero…!



… but how to sample $Z_{t+1} \sim$ posterior?

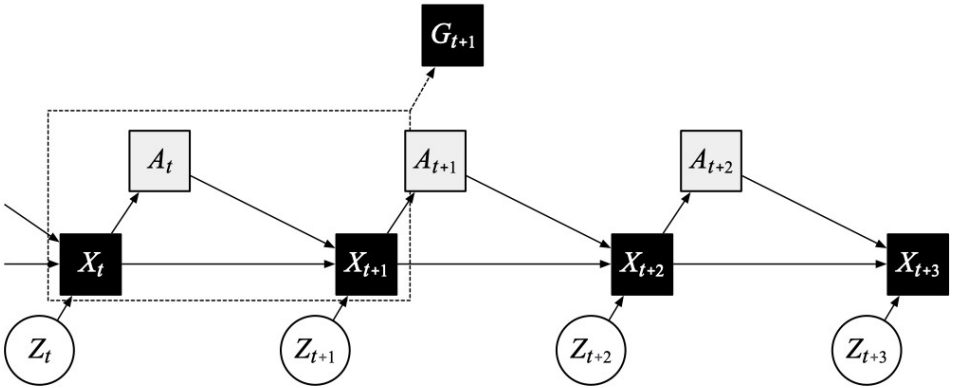
Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



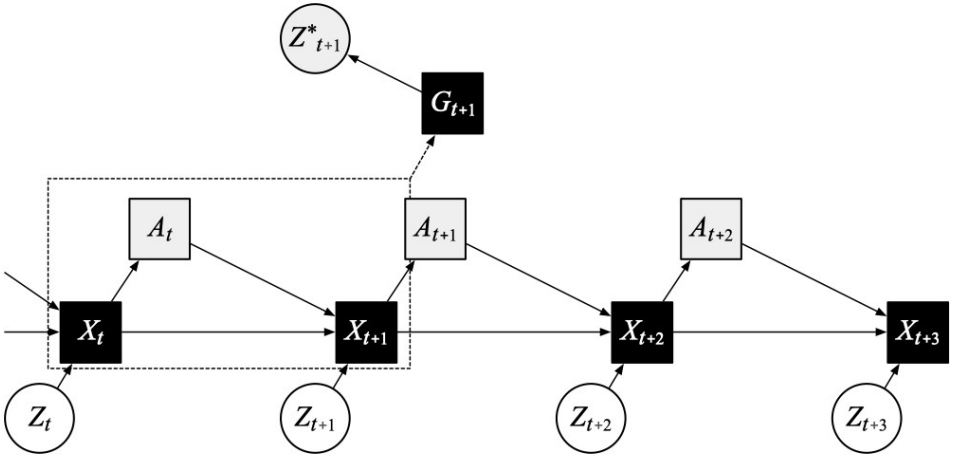
Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



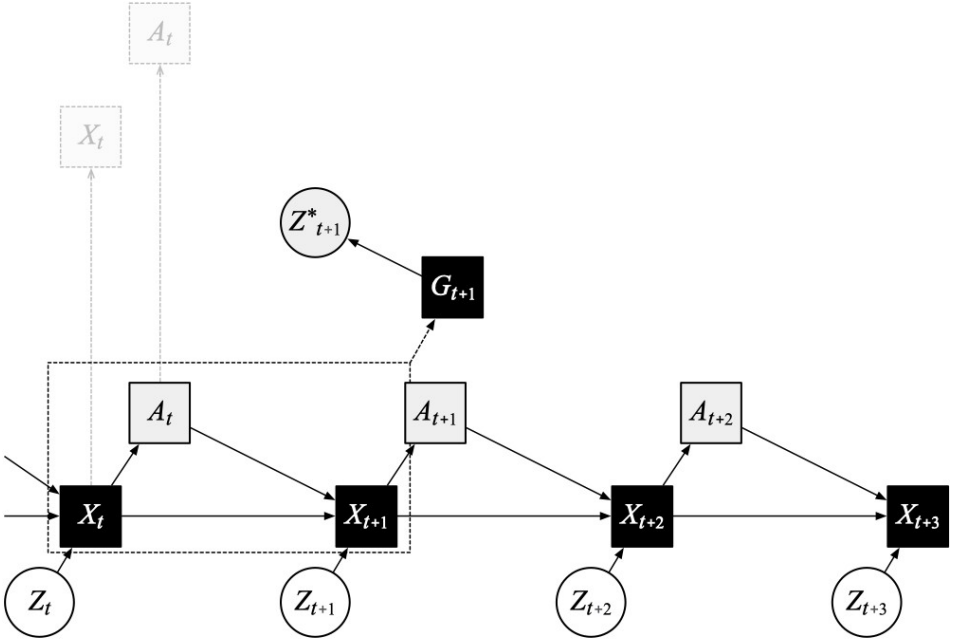
Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



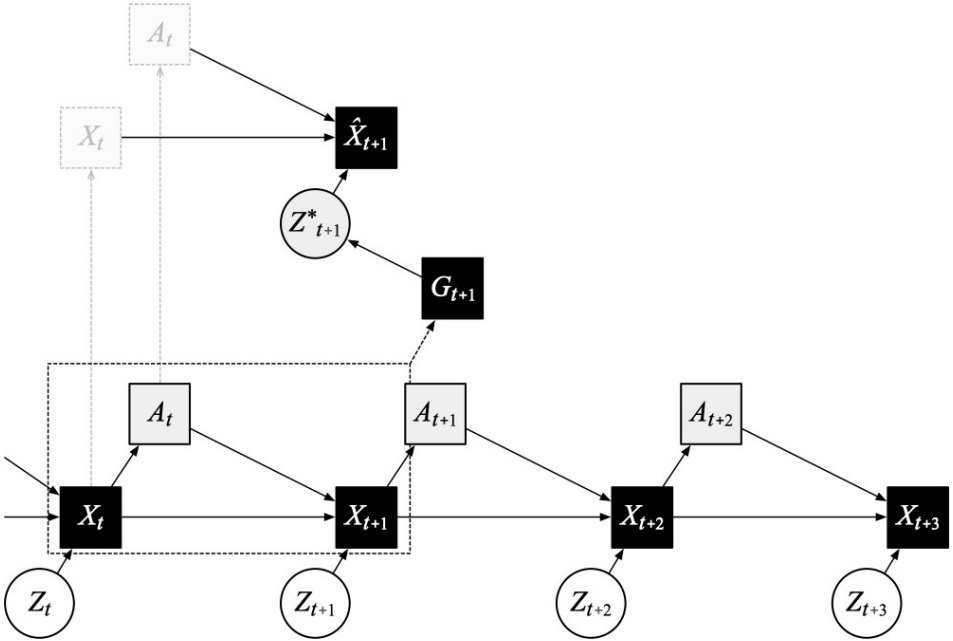
Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



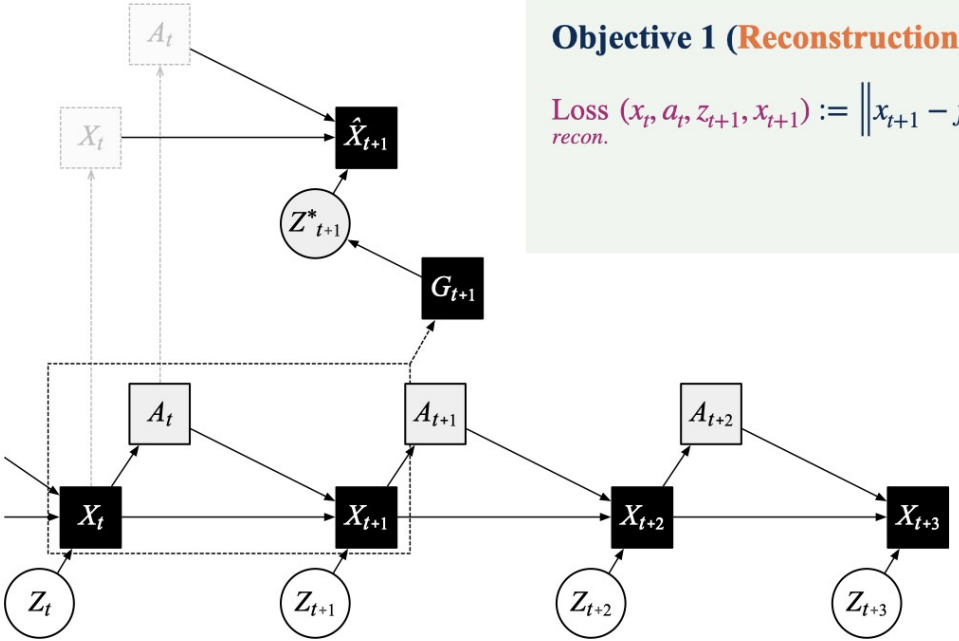
Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.

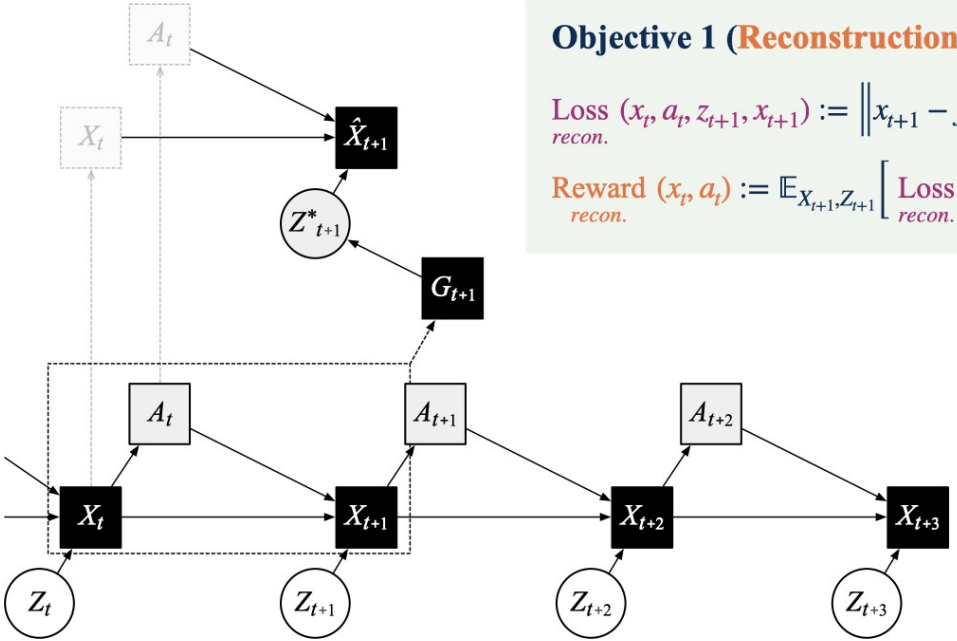


Objective 1 (Reconstruction)

$$\underset{\textit{recon.}}{\text{Loss}}\ (x_t, a_t, z_{t+1}, x_{t+1}) := \left\| x_{t+1} - \hat{f}(x_t, a_t, z_{t+1}) \right\|_2^2$$

Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



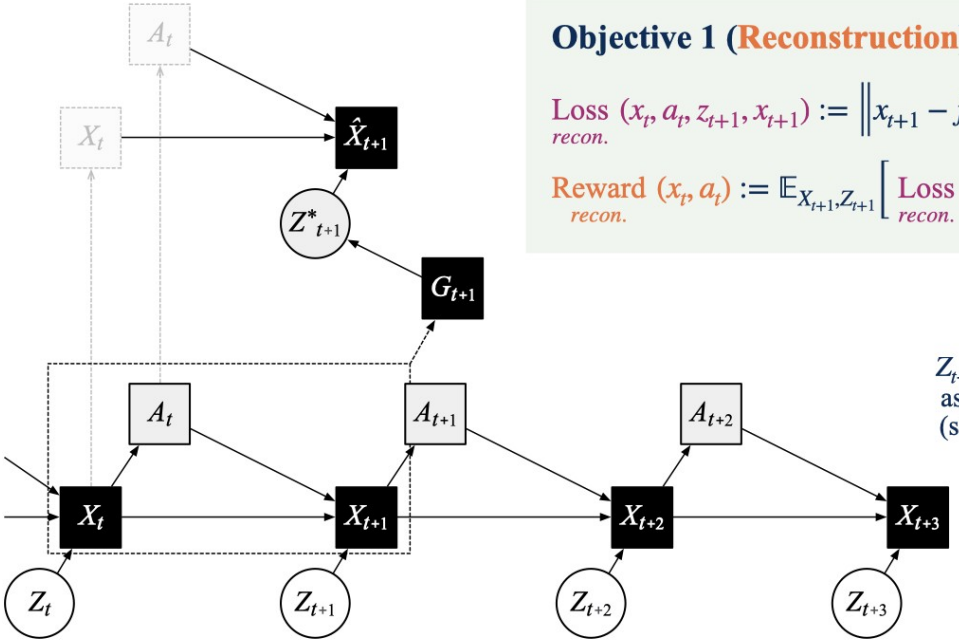
Objective 1 (Reconstruction)

$$\underset{\text{recon.}}{\text{Loss}}\,(x_t, a_t, z_{t+1}, x_{t+1}) := \left\| x_{t+1} - \hat{f}(x_t, a_t, z_{t+1}) \right\|_2^2$$

$$\underset{\text{recon.}}{\text{Reward}}\,(x_t, a_t) := \mathbb{E}_{X_{t+1}, Z_{t+1}}\left[\underset{\text{recon.}}{\text{Loss}}\,(x_t, a_t, Z_{t+1}, X_{t+1}) \right]$$

Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



Objective 1 (Reconstruction)

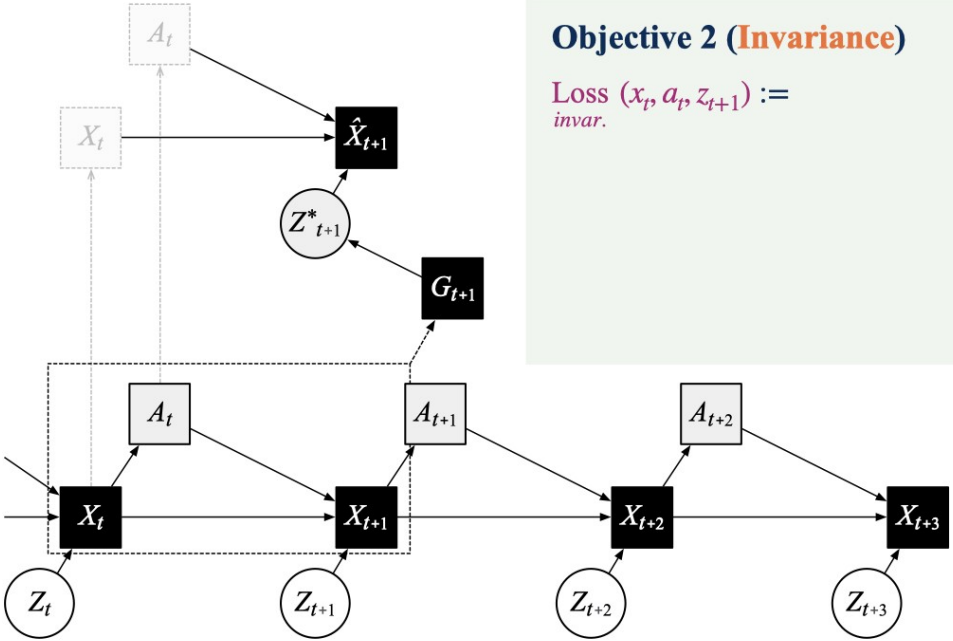
$$\underset{\text{recon.}}{\text{Loss}}\,(x_t, a_t, z_{t+1}, x_{t+1}) := \left\| x_{t+1} - \hat{f}(x_t, a_t, z_{t+1}) \right\|_2^2$$

$$\underset{\text{recon.}}{\text{Reward}}\,(x_t, a_t) := \mathbb{E}_{X_{t+1}, Z_{t+1}}\left[\underset{\text{recon.}}{\text{Loss}}\,(x_t, a_t, Z_{t+1}, X_{t+1}) \right]$$

Z_{t+1} should capture *at least* all aspects that are unpredictable (so we don't reward the agent for *irreducible* error).

Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.

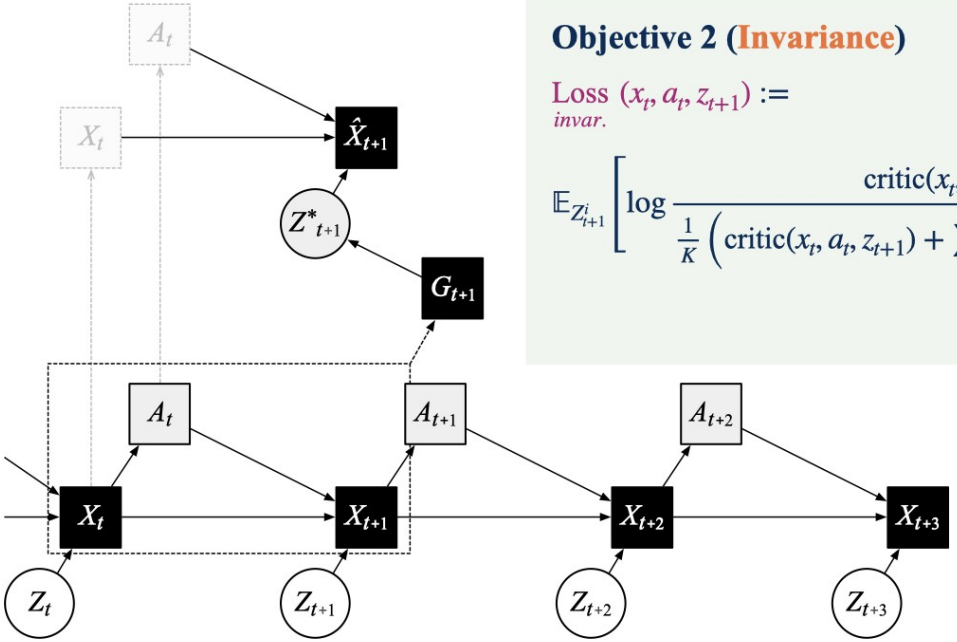


Objective 2 (Invariance)

Loss (x_t, a_t, z_{t+1}) :=
invar.

Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.

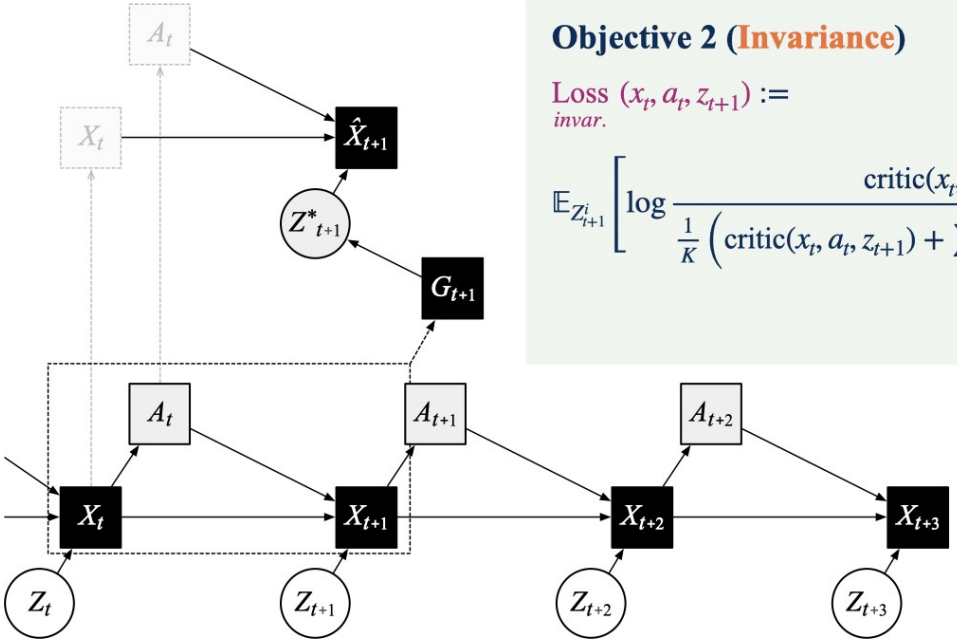


Objective 2 (Invariance)

$$\text{Loss}_{\text{invar.}}(x_t, a_t, z_{t+1}) := \mathbb{E}_{Z^i_{t+1}}\left[\log \frac{\text{critic}(x_t, a_t, z_{t+1})}{\frac{1}{K}\left(\text{critic}(x_t, a_t, z_{t+1}) + \sum_{i=1}^{K-1} \text{critic}(x_t, a_t, Z^i_{t+1})\right)}\right]$$

Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



Objective 2 (Invariance)

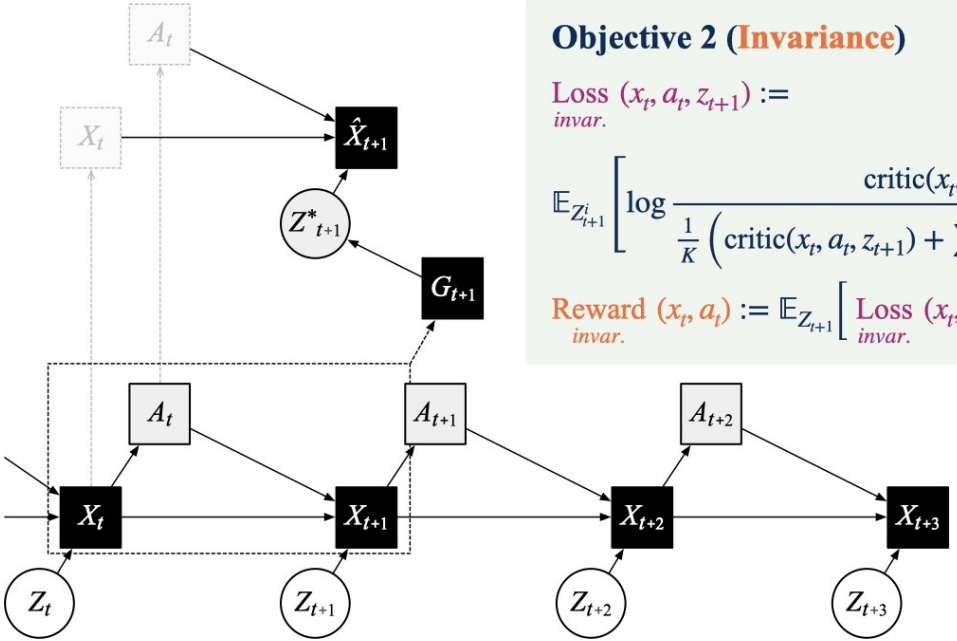
positive sample

negative sample

$$\text{Loss}_{\text{invar.}}(x_t, a_t, z_{t+1}) := \mathbb{E}_{Z^i_{t+1}}\left[\log \frac{\text{critic}(x_t, a_t, z_{t+1})}{\frac{1}{K}\left(\text{critic}(x_t, a_t, z_{t+1}) + \sum_{i=1}^{K-1} \text{critic}(x_t, a_t, Z^i_{t+1})\right)}\right]$$

Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



Objective 2 (Invariance)

positive sample

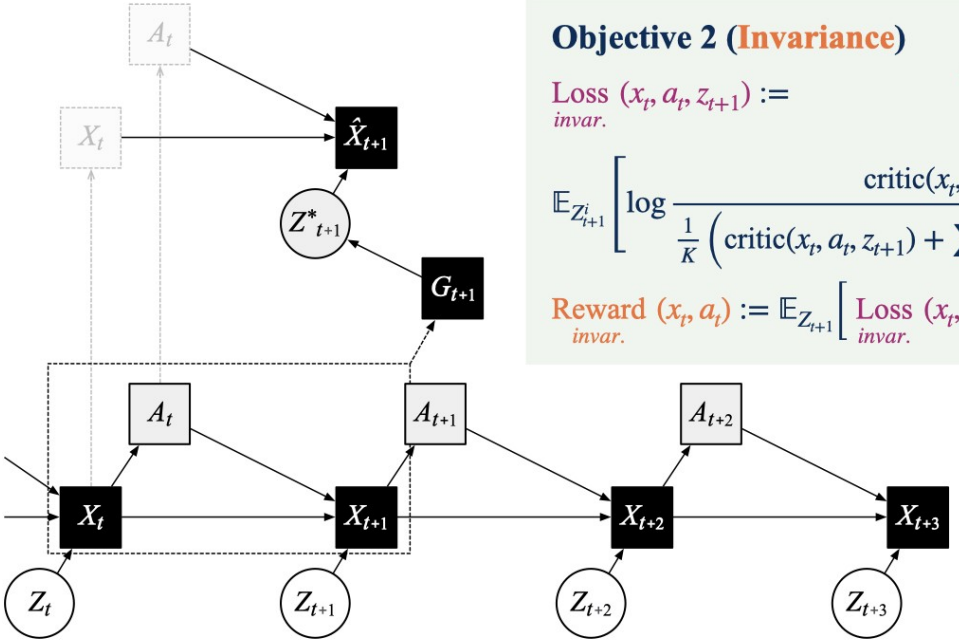
negative sample

$$\text{Loss}_{\text{invar.}}(x_t, a_t, z_{t+1}) := \mathbb{E}_{Z^i_{t+1}}\left[\log \frac{\text{critic}(x_t, a_t, z_{t+1})}{\frac{1}{K}\left(\text{critic}(x_t, a_t, z_{t+1}) + \sum_{i=1}^{K-1} \text{critic}(x_t, a_t, Z^i_{t+1})\right)}\right]$$

$$\text{Reward}_{\text{invar.}}(x_t, a_t) := \mathbb{E}_{Z_{t+1}}\left[\text{Loss}_{\text{invar.}}(x_t, a_t, Z_{t+1})\right]$$

Curiosity in Hindsight — Hindsight Representations

Learn a **Reconstructor** $\hat{f}(X_t, A_t, Z_{t+1})$ and a **Generator** $\hat{p}(Z_{t+1} \mid X_t, A_t, X_{t+1})$.



Objective 2 (Invariance)

positive sample

negative sample

$$\text{Loss}_{\text{invar.}}(x_t, a_t, z_{t+1}) := \mathbb{E}_{Z^i_{t+1}}\left[\log \frac{\text{critic}(x_t, a_t, z_{t+1})}{\frac{1}{K}\left(\text{critic}(x_t, a_t, z_{t+1}) + \sum_{i=1}^{K-1} \text{critic}(x_t, a_t, Z^i_{t+1})\right)}\right]$$

$$\text{Reward}_{\text{invar.}}(x_t, a_t) := \mathbb{E}_{Z_{t+1}}\left[\text{Loss}_{\text{invar.}}(x_t, a_t, Z_{t+1})\right]$$

Z_{t+1} should capture *at most* all aspects that are unpredictable (so we do reward the agent for *reducible* error).

Curiosity in Hindsight — Optimistic Exploration

Curiosity in Hindsight — Optimistic Exploration

Definition 1 (Curiosity)

$$\underset{\text{pred.}}{\text{Reward}}\ (x_t, a_t) := \mathbb{E}_{X_{t+1}} \left\| X_{t+1} - \hat{x}_{t+1} \right\|_2^2$$

The agent performs

$$\underset{\pi}{\text{(policy) maximize}}\ \ \underset{\hat{\tau}}{\text{(model) min}}\ \ \mathbb{E}_{X_t, A_t}\left[\underset{\text{pred.}}{\text{Reward}}\ (X_t, A_t) \right]$$

Curiosity in Hindsight — Optimistic Exploration

Definition 1 (Curiosity)

$$\underset{\textit{pred.}}{\text{Reward}}\ (x_t, a_t) := \mathbb{E}_{X_{t+1}} \left\| X_{t+1} - \hat{x}_{t+1} \right\|_2^2$$

The agent performs

$$\underset{\pi}{\text{(policy) maximize}}\ \ \underset{\hat{\tau}}{\text{(model) min}}\ \ \mathbb{E}_{X_t, A_t}\left[\underset{\textit{pred.}}{\text{Reward}}\ (X_t, A_t) \right]$$

Definition 2 (Curiosity in Hindsight)

$$\underset{\textit{hindsight}}{\text{Reward}}\ (x_t, a_t) := \frac{1}{\lambda} \underset{\textit{recon.}}{\text{Reward}}\ (x_t, a_t) + \lim_{K \to \infty} \underset{\textit{invar.}}{\text{Reward}}\ (x_t, a_t)$$

The agent performs

$$\underset{\pi}{\text{(policy) maximize}}\ \ \underset{\hat{p}, \hat{\tau}}{\text{(model) min}}\ \ \mathbb{E}_{X_t, A_t}\left[\underset{\textit{hindsight}}{\text{Reward}}\ (x_t, a_t) \right]$$

Curiosity in Hindsight — Optimistic Exploration

Optimistic Exploration: For an optimal critic and "sufficiently small"…

$$\underset{\textit{hindsight}}{\text{Reward}}\ (x_t, a_t) \geq \text{Divergence}\Big(\underset{\text{(reality)}}{\tau(X_{t+1} \mid x_t, a_t)} \,\Big\|\, \underset{\text{(model)}}{\hat{\tau}(X_{t+1} \mid x_t, a_t)} \Big)$$

…and converges to zero in the limit.

Definition 2 (Curiosity in Hindsight)

$$\underset{\textit{hindsight}}{\text{Reward}}\ (x_t, a_t) := \frac{1}{\lambda} \underset{\textit{recon.}}{\text{Reward}}\ (x_t, a_t) + \lim_{K \to \infty} \underset{\textit{invar.}}{\text{Reward}}\ (x_t, a_t)$$

The agent performs

$$\underset{\pi}{\text{(policy) maximize}}\ \ \underset{\hat{p}, \hat{\tau}}{\text{(model) min}}\ \ \mathbb{E}_{X_t, A_t}\Big[\underset{\textit{hindsight}}{\text{Reward}}\ (x_t, a_t) \Big]$$

Overview

1. **Introduction**: Curiosity-driven Exploration
2. **Motivation**: Stochastic Environments
3. **Curiosity in Hindsight**
4. **Practical Framework**
5. **Experiments**
6. **Future Directions**

Practical Framework

In Practice: (1) batch size $K < \infty$, (2) critic not fully optimized, (3) λ is a hyperparameter.

Practical Framework

In Practice: (1) batch size $K < \infty$, (2) critic not fully optimized, (3) λ is a hyperparameter.

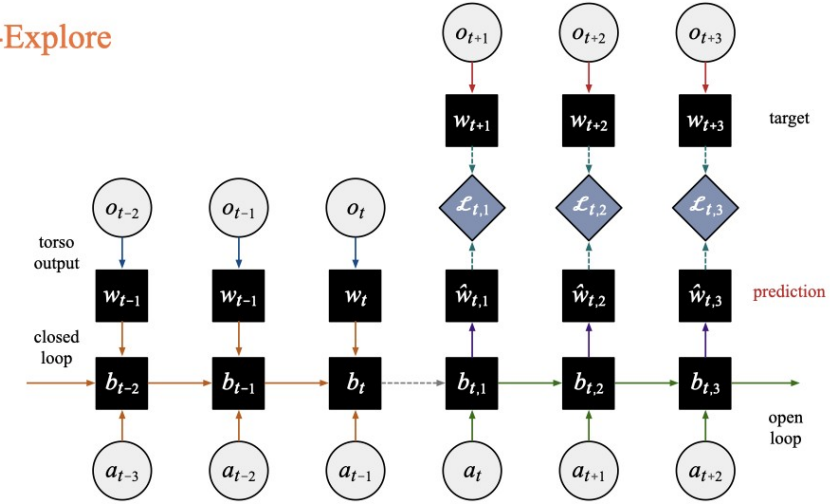
Overall, simple **drop-in modification** on top of any curiosity-driven exploration method.

Practical Framework

In Practice: (1) batch size $K < \infty$, (2) critic not fully optimized, (3) λ is a hyperparameter.

Overall, simple **drop-in modification** on top of any curiosity-driven exploration method.

Example: BYOL-Explore

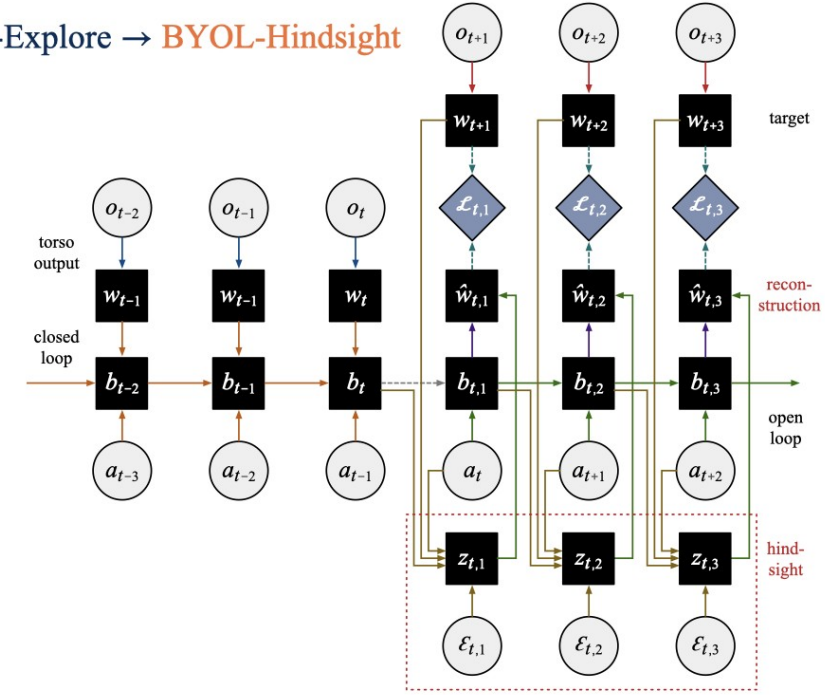


Practical Framework

In Practice: (1) batch size $K < \infty$, (2) critic not fully optimized, (3) λ is a hyperparameter.

Overall, simple **drop-in modification** on top of any curiosity-driven exploration method.

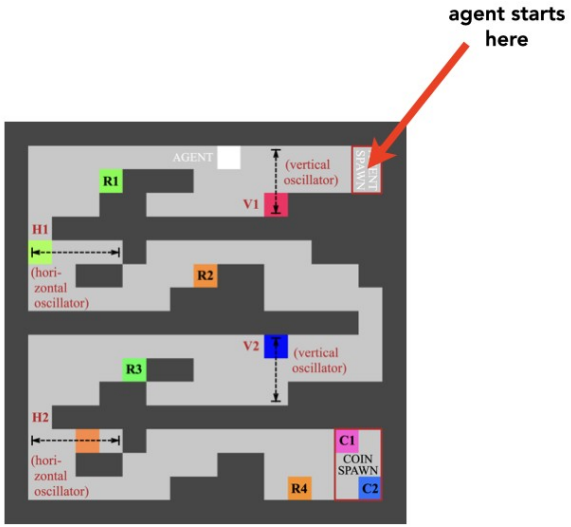
Example: BYOL-Explore → BYOL-Hindsight



Overview

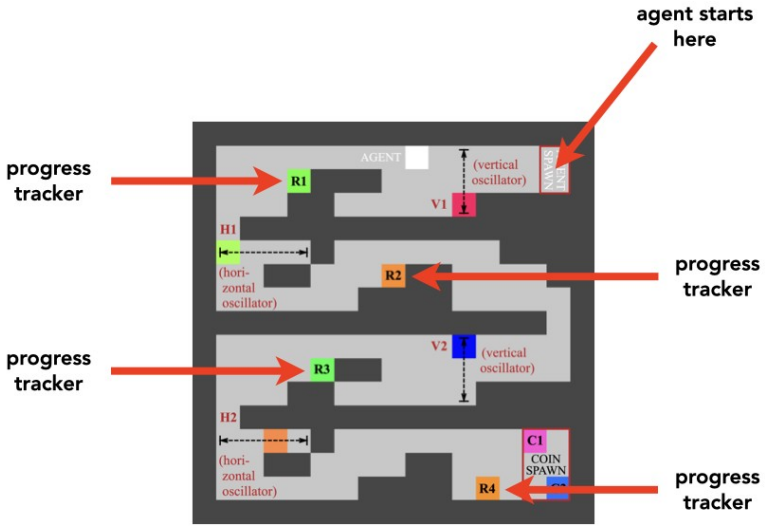
1. **Introduction**: Curiosity-driven Exploration
2. **Motivation**: Stochastic Environments
3. **Curiosity in Hindsight**
4. **Practical Framework**
5. **Experiments**
6. **Future Directions**

Stress Test — Robustness to Different Stochasticities



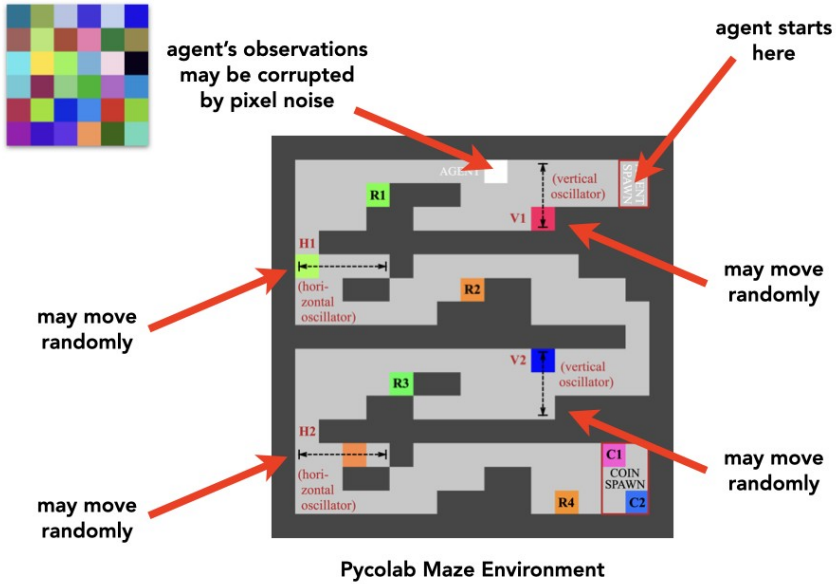
Pycolab Maze Environment

Stress Test — Robustness to Different Stochasticities



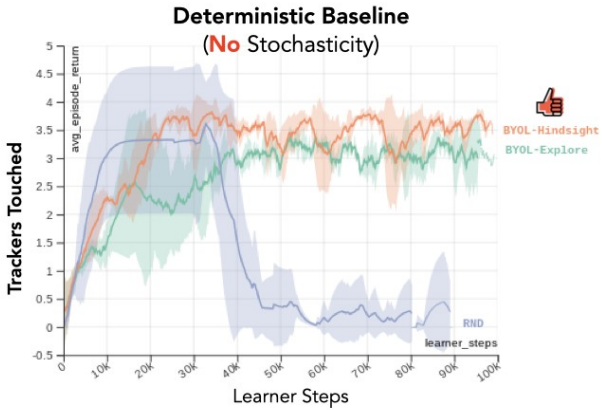
Pycolab Maze Environment

Stress Test — Robustness to Different Stochasticities

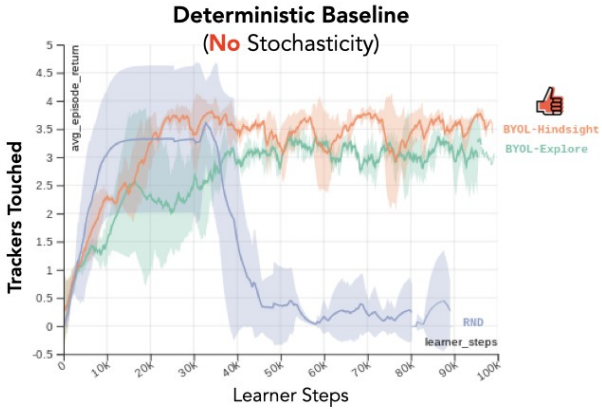


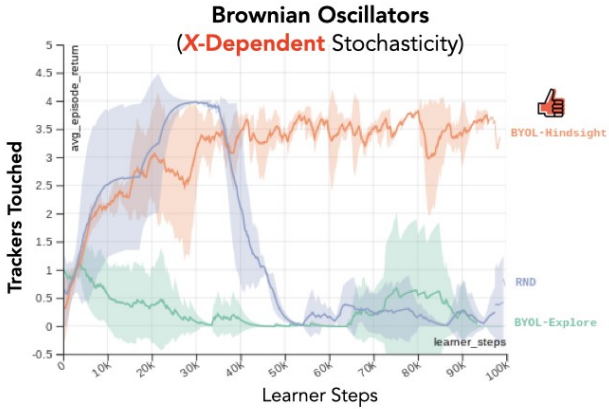
Pycolab Maze Environment

Stress Test — Robustness to Different Stochasticities



Stress Test — Robustness to Different Stochasticities

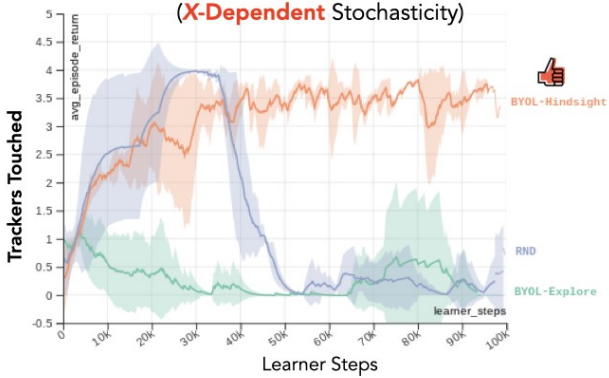




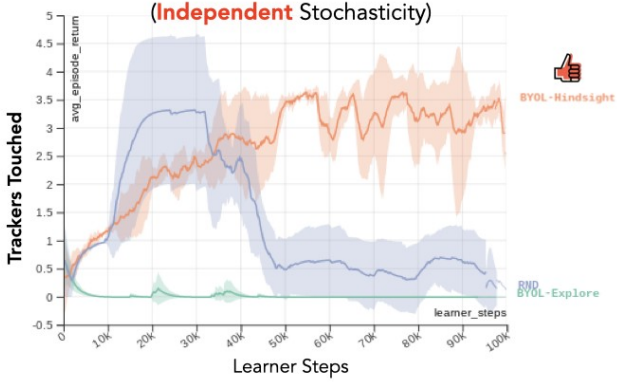
Stress Test — Robustness to Different Stochasticities

Deterministic Baseline
(**No** Stochasticity)

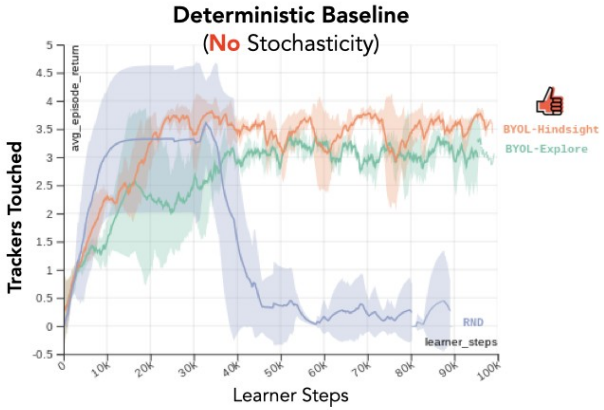
Brownian Oscillators
(**X-Dependent** Stochasticity)

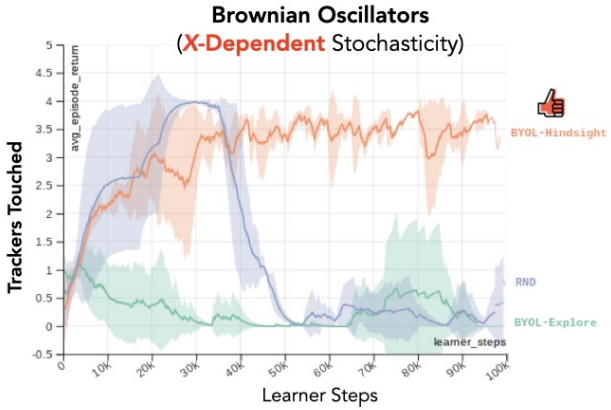


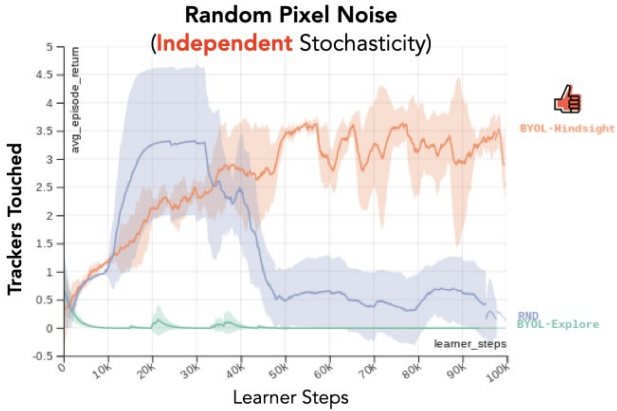
Random Pixel Noise
(**Independent** Stochasticity)

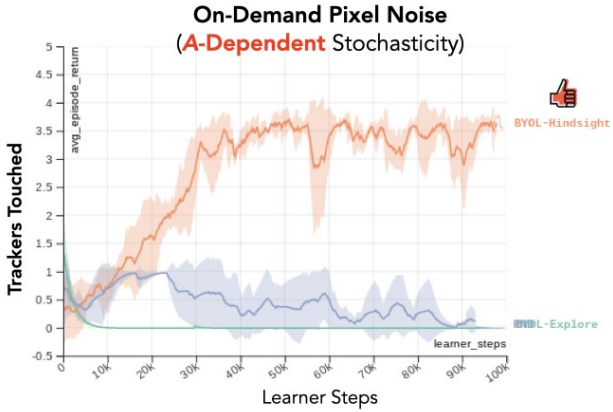


Stress Test — Robustness to Different Stochasticities

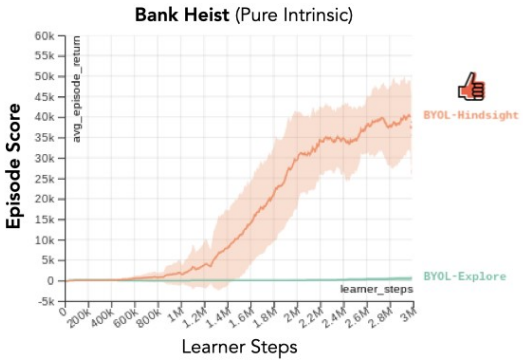


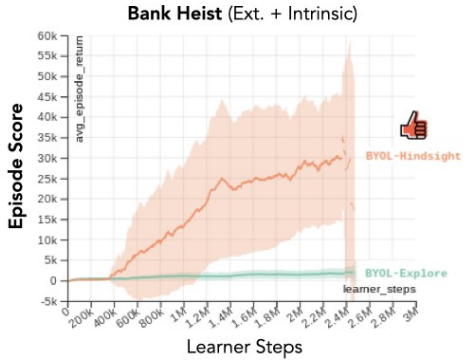




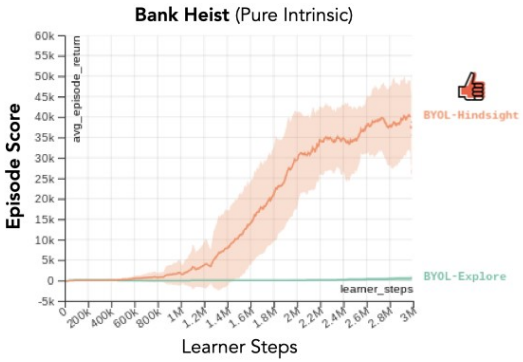


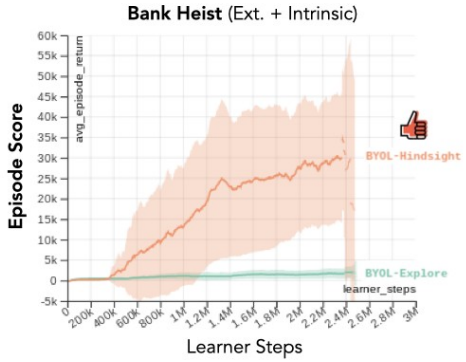
Atari — Hard Exploration (Bank Heist)

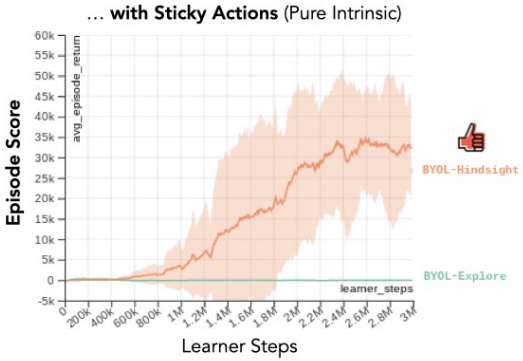


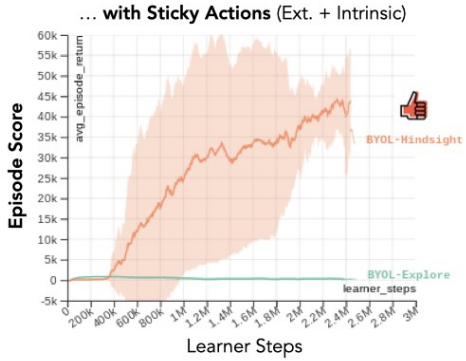


Atari — Hard Exploration (Bank Heist)

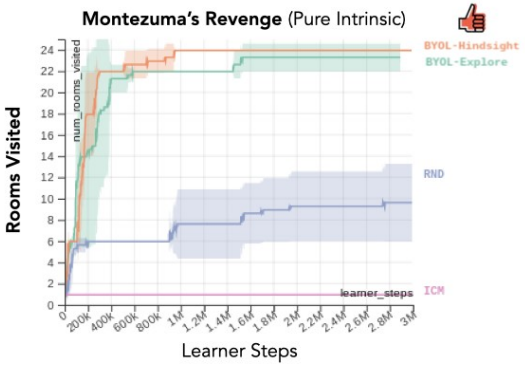


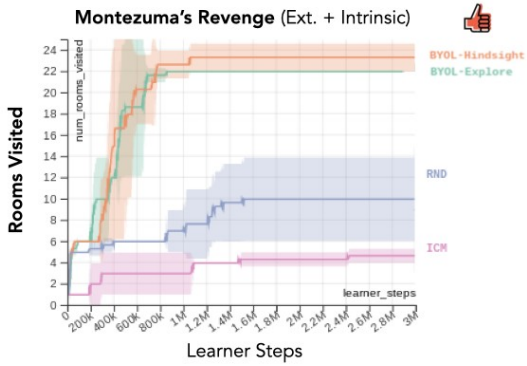




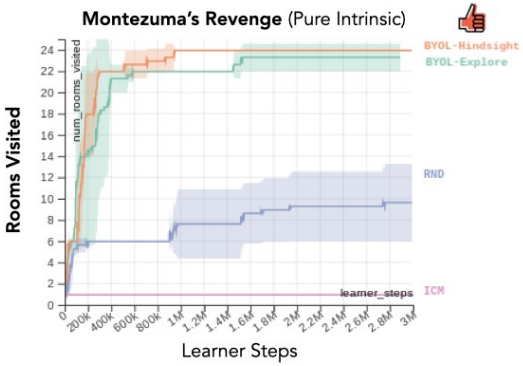


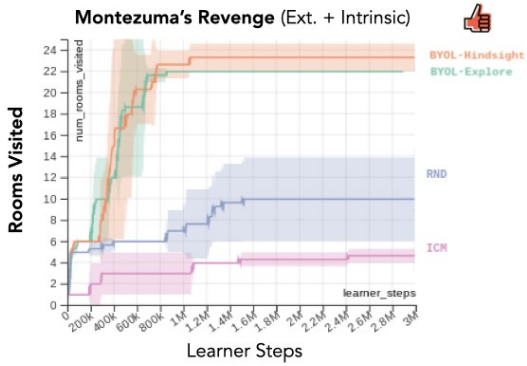
Atari — Hard Exploration (Montezuma's Revenge)

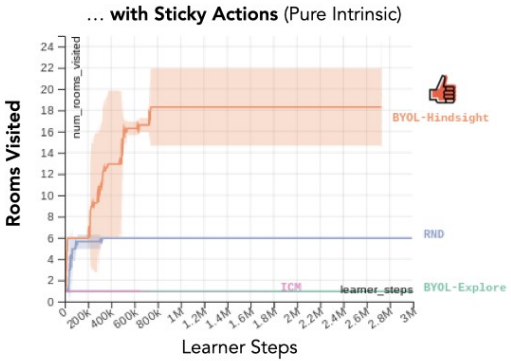


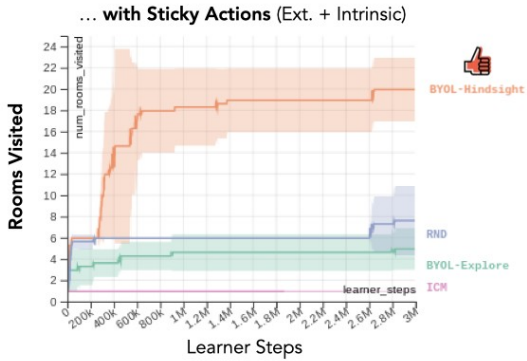


Atari — Hard Exploration (Montezuma's Revenge)

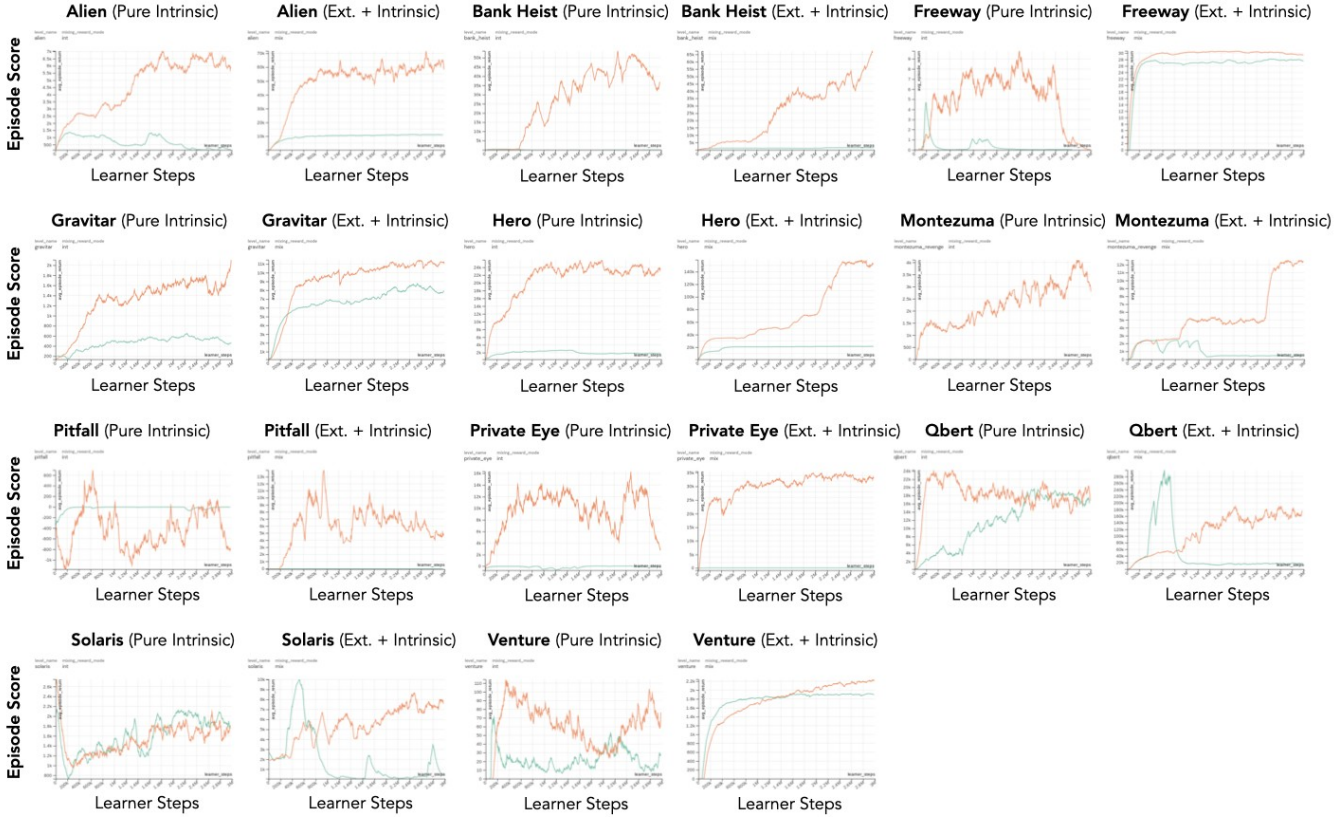








Atari — Hard Exploration (with Sticky Actions)



Overview

1. **Introduction**: Curiosity-driven Exploration
2. **Motivation**: Stochastic Environments
3. **Curiosity in Hindsight**
4. **Practical Framework**
5. **Experiments**
6. **Future Directions**

Recap + Future Directions

How to **refine** curiosity?	How to **measure** curiosity?	How to **implement** curiosity?
(Learnable) Epistemic Knowledge vs. (Unlearnable) Aleatoric Variation	(Future-summarizing) Hindsight Representations vs. (History-summarizing) Context Representations	(Reducible) Intrinsic Rewards vs. (Irreducible) Stochastic Traps

Exploration

Inconsistency as a signal for exploration

Credit Assignment

Different levels of hindsight-conditioning

Multiple Agents

Other agents as “stochasticity”