Using adaptive intrinsic motivation in RL to model learning across development

Kai Sandbrink[*]
University of Oxford, Oxford
United Kingdom

Brian Christian
University of Oxford, Oxford
United Kingdom

Linas Nasvytis
Harvard University, Cambridge
United States

Christian Schroeder de Witt
University of Oxford, Oxford
United Kingdom

Patrick Butlin
University of Oxford, Oxford
United Kingdom

Abstract

Reinforcement learning is a powerful model of animal learning in brief, controlled experimental conditions, but does not readily explain the development of behavior over an animal's whole lifetime. In this paper, we describe a framework to address this shortcoming by introducing the single-life reinforcement learning setting to cognitive science. We construct an agent with two learning systems: an extrinsic learner that learns within a single lifetime, and an intrinsic learner that learns across lifetimes, equipping the agent with intrinsic motivation. We show that this model outperforms heuristic benchmarks and recapitulates a transition from exploratory to habit-driven behavior, while allowing the agent to learn an interpretable value function. We formulate a precise definition of intrinsic motivation and discuss the philosophical implications of using reinforcement learning as a model of behavior in the real world.

1 Introduction

Reinforcement learning (RL), in which an agent learns to optimize expected rewards by interacting with an environment, is both a powerful model of learning in cognitive science [1, 2, 3], and a successful training paradigm machine learning [4, 5]. In both cases, however, the reward signal offered by the environment is too sparse to itself fully describe learning, a problem called the “sparse rewards problem." To address this, handcrafted heuristics such as count-based intrinsic motivation in machine learning [6] or novelty- and stochasticity-seeking behavior in humans [7, 8] are frequently used to supplement extrinsic rewards. Hand-crafting intrinsic motivation and intrinsic rewards in machine learning, however, can lead to unpredictable agent behavior [9]. In reality, biological agents have no access to hand-crafted intrinsic motivation and reward functions, and must construct their own sense of what is rewarding [10].

In this paper, we introduce deep RL networks that use meta-RL to learn intrinsic motivation. Unlike other recent suggestions that learn an agent-internal reward function [11], we focus specifically on meta-learning intrinsic motivation (defined as bonuses to actions before action selection) while continuing to use an environmentally-determined reward function. This method allows us to focus specifically on the role of meta-learning in determining our algorithms of exploration and action. We recapitulate changes in intrinsic motivation that could capture the adaptation of “exploratory hyperparameters" across development [12]. This method could in theory be combined with other approaches to address shortcomings of hand-crafted models, such as meta-learning a time-dependent

[*] Corresponding author, kai.sandbrink@lmh.ox.ac.uk

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

policy or an intrinsic reward function [13, 14]. However, meta-learning intrinsic motivation has several key advantages: First, it allows learning a value function that represents the true extrinsic rewards in the environment. Second, it makes explicit in which directions agents are driven by extrinsic reward, and when the motivation is intrinsic. Finally, it reduces the amount of assumptions needed in training learning agents on a task *de novo*, allowing principled studies of learning dynamics.

2 Methods

2.1 Single-life reinforcement learning

We model the learning of extrinsic rewards and adaptation of intrinsic motivation as taking place over a single life. The defining characteristic of the single-life reinforcement learning (SLRL) setting is that the agent is given a single "life" (i.e. one long episode) over which to accumulate rewards [15].

The agent interacts with a Markov decision process (MDP; [16] $M_{\text{life}} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, p_0, \gamma)$ sampled from $\mathcal{M}_{\text{evol}}$. Its goal is to maximize $G^{\text{life}} = \sum_{t=0}^{h} \gamma^t \mathcal{R}(s_t)$ over the course of a single episode, which may be infinitely long but normally ends with a terminal state in the MDP. The agent's trajectory over the episode is called the lifetime trajectory τ and follows the distribution $p_\eta(\tau|\theta_0) = p(s_0)\prod_{t=0}^{T-1} \pi_{\theta_t,\psi}(a_t|s_t)p(r_{t+1}, s_{t+1}|s_t, a_t)$, where $\theta_t = f(\theta_{t-1}, \psi)$ are the policy parameters of the extrinsic learner.

2.2 Optimal intrinsic motivation

We conceptualize intrinsic motivation generally as motivation which is trained using extrinsic rewards across lifespans, but is not based on any extrinsic rewards during a single lifespan. In analogy to previous work on intrinsic rewards [13, 17], we define the *Optimal Intrinsic Motivation Problem* as learning the intrinsic motivation that maximizes the expected value of the lifetime return G^{life} obtained by the combined learning agent within a lifetime. Unlike previous work, however, we are learning a separate intrinsic motivation policy which only takes into account state-action trajectories and is combined convexly with an extrinsically-generated policy.

We address this problem by meta-learning across lifetimes. Meta-learning occurs over a set $\mathcal{M}_{\text{evol}}$ of Markov decision processes (MDPs) from which we sample according to a distribution $\rho_{\mathcal{M}_{\text{evol}}} : \mathcal{M} \rightarrow \mathbb{R}_+$ at each new lifetime [18, 19]. The objective function of this meta-learning timescale is

$$J = \mathbf{E}_{\theta_0 \sim \Theta, M_{\text{life}} \sim p(\mathcal{M}_{\text{evol}})} \left[\mathbf{E}_{\tau \sim p_\psi(\tau|\theta_0)} \left[G^{\text{life}} \right] \right] \tag{1}$$

where Θ is an initial policy distribution of the extrinsic learner, ψ are the parameters of the generative model for intrinsic motivation, and τ is a single-life history of the combined agent. We train the agent on stationary bandit tasks (task 1), fixed reward structures (task 2), and volatile environments (task 3). $\mathcal{M}_{\text{evol}}$ represents all problems in the considered distributions, whereas $\mathcal{M}_{\text{life}}$ are the instances sampled uniformly from these sets. For tasks 1 and 2, the observation consists solely of the action selected by the agent on the previous turn; for task 3, in these simulations, the observation additionally includes reward feedback from the sampled arm to alert the agent to a change.

2.3 A reward-learning, adaptive-intrinsic-motivation agent

We build an agent that is composed of two components, an extrinsic learner that begins every episode without prior knowledge about the environment and a meta-learning intrinsic motivation learner. We thus operationalize intrinsic motivation as motivation which is not based on rewards information, even if it is trained by extrinsic rewards across episodes.

The *extrinsic learner* has at its objective to maximize the episodic return G^{life}. Its values or parameters are updated in an online manner after every timestep or at least several times within an episode. In our experiments, the extrinsic learner is implemented as a tabular Q-learning system [20] initialized to 0 with learning rate η. For stationary tasks, $\eta = 1/N(a)$, where $N(a)$ is the number of times an action a was chosen. With this value, the Q-values track the means of the bandits across observations. For non-stationary tasks, we set a non-decaying learning rate $\eta = 0.1$.

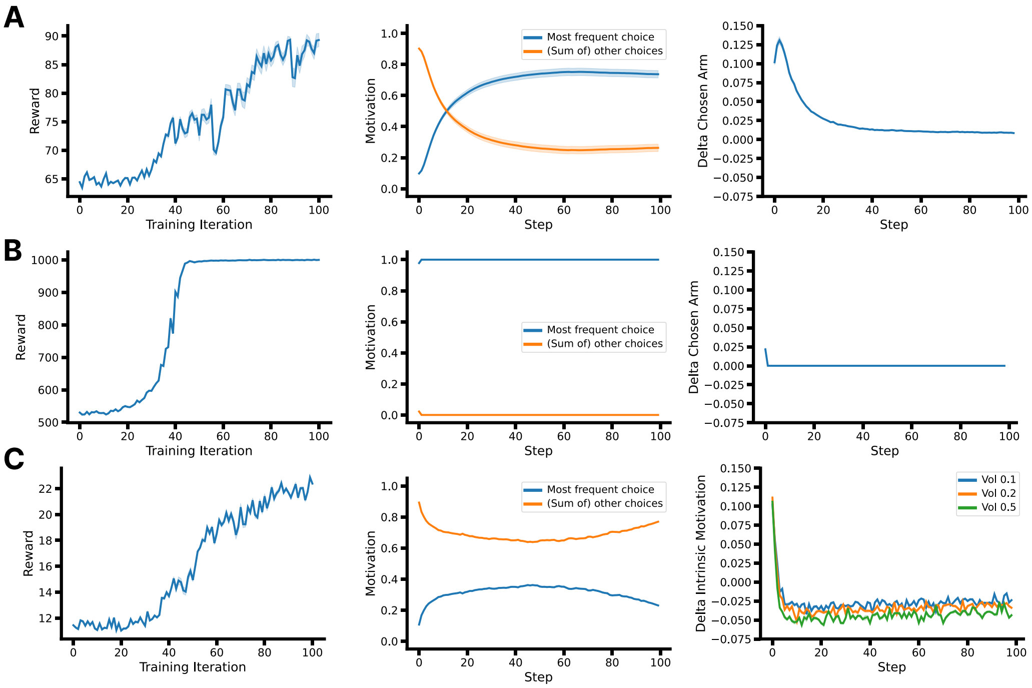

Figure 1: Behaviour of the intrinsic-motivation learning module on variants of the standard 10-armed bandit testbed. **A.** Behavior on the stationary 10-armed bandit task. (*left*) Learning curves as measured on the training distribution for each episode over 1000 instantiations for tasks 1 and 3 and 100 for task 2. (*middle*) Intrinsic motivation attributed to the dominant arm over the course of the whole experiment as sampled over 1000 instantiations for all tasks. (*right*) How much the intrinsic motivation for a given arm is updated after that arm is selected as sampled over 1000 instantiations for all tasks. **B.** Same as A, but for the constant 10-armed bandit task in which the distribution of bandits remains the same across different episodes. **C** Same as B, but for non-stationary bandit tasks in which the amount of volatility changes between different episodes.

The *intrinsic learner* has the objective given by Equation 1, and outputs an intrinsic policy over actions based on state-action trajectories. Its parameters are updated after exposure to a batch of different lifetimes to ensure that it learns parameters that are useful across different MDPs drawn from $\mathcal{M}_{\text{evol}}$. We construct this agent as a meta-reinforcement learner trained as in Wang et al. [18]. Except in the task where we test for responses to volatility, the intrinsic learner only receives action history information as input (and not reward information). We implement the intrinsic learner as a network composed of a Long Short-Term Memory (LSTM; [21]) layer of 64 units followed by a softmax output layer for action selection. We use the REINFORCE algorithm [22] to train the network. In all simulations, we train the network for 500,000 episodes, with annealed entropy regularization to 0 over the course of the first 250,000 episodes.

Both learners output policies for every timestep. These are combined convexly into one global policy based on a mixture weight $\alpha \in [0, 1]$, such that $\pi_{\text{agent}} = (1 - \alpha) \times \pi_{\text{extr}} + \alpha \times \pi_{\text{intr}}$. We intentionally set a high value of $\alpha = 0.5$ to study the impact of the intrinsic motivation on behavior.

3 Results

3.1 Learned intrinsic motivation in the ten-armed bandit testbed

First, we model performance on the standard ten-armed bandit testbed [4] over episodes of 100 steps. We train the meta-learner over 500,000 episodes. The payout magnitude of each action is sampled from a standard normal distribution $\mathcal{N}(0, 1)$ at the beginning of each episode. Across five model instantiations, after training, the model reaches an average performance of 87.7 ± 25.9 (mean ± SEM over models, see Figure 1A). We compare the performance of our system with other models of intrinsic motivation. A 0.5-greedy system (that has the same strength of intrinsic motivation) has an average reward of 59.9 ± 3.7 (mean ± SEM over 100 test episodes) on the same system. Upper Confidence Bound (UCB; 23) of the same strength has a performance of 72.9 ± 3.8 (mean ± SEM over 100 test episodes). This illustrates that the learned intrinsic motivation in the system significantly outperforms handcrafted heuristics (one-sample t-test comparing average performance for each of the different model instantiations with average performance ε-greedy: t(4)=17.08, p=3.4e-5, UCB: t(4)=17.06, p=3.5e-5).

3.2 Evolutionarily-transmitted knowledge

Second, we consider situations where the distributions in payout magnitude across bandits has a fixed structure across episodes. We sample the payout of arm 1 from the higher distribution $\mathcal{N}(10, 1)$. In this case, the model achieves an average performance of 999.9 ± 0.36 (mean ± SEM over models). Figure 1B illustrates that the intrinsic motivation system knows which arm to incentivize from the first step of the episode. This result highlights that the system is capable of modelling instinctive responses such as fear and innate attraction using our definition of intrinsic motivation. In contrast, the ε-greedy system achieves rewards of 532.7 ± 9.1 (mean ± SEM over 100 test episodes) in this case. The UCB system achieves average rewards of 908.2 ± 3.2 (mean ± SEM over 100 test episodes). Both are significantly worse than the system with adaptive intrinsic rewards (one-sample t-test comparing average performance for each of the different model instantiations with average performance ε-greedy: $t(4) = 1167.4$, $p = 1.6e - 12$, UCB: $t(4) = 229.1$, $p = 1.1e - 9$).

3.3 Within-lifetime adaptation of exploratory policies

Finally, we show that when given access to extrinsic change signals such as reward feedback, the intrinsic motivation adapts to early experience. Figure 1C illustrates how the intrinsic motivation supplied to the extrinsic learner differs across implementations of the bandit task, where the arms have a 10%, 20%, and 50% chance of being redrawn from the base distribution $\mathcal{N}(0, 1)$ between different trials in the same episode (volatility).

Agents were trained on volatility levels in the intervals $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$. Statistical analyses reveal significant differences in performance adaptation between our model and both the UCB and ε-greedy across volatility settings. For UCB, one-sample t-tests yield t(4)=7.0, p=0.0011 at 10% volatility, t(4)=-2.0, p=0.944 at 20% volatility, and t(4)=18.0, p=3.1e-5 at 50% volatility. For ε-greedy, the tests yield t(4)=46.0, p=6.8e-7 at 10% volatility, t(4)=23.0, p=9.9e-6 at 20% volatility, and t(4)=-18.0, p=1.0 at 50% volatility, underscoring our model's enhanced capability to modulate exploration in response to environmental volatility shifts.

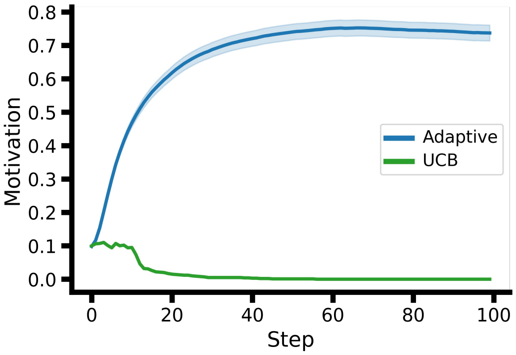

Figure 2: Comparison of (*blue*) meta-learned intrinsic motivation terms with (*green*) the hand-crafted heuristic upper-confidence bound (UCB) showing intrinsic motivation attributed to the dominant arm over an episode. Since this arm is chosen increasingly frequently by the agent, it is less likely to selected by UCB. The adaptive intrinsic motivation meanwhile changes from encouraging exploration early in the episode to habit-driven learning later on.

3.3.1 Comparing learned intrinsic motivation function with hand-crafted heuristics

The meta-learned intrinsic motivation follows a smooth transition from encouraging exploration (intrinsic motivation is spread across the ten arms) to exploitation (intrinsic motivation is concentrated on one particular arm). In contrast, hand-crafted heuristics favor exploration even after it has stopped being beneficial (Figure 2).

4 Conclusion

The primary aim of this paper is to show how
reinforcement learning can be used as a model of learning across development in a single lifetime. We show that we can model the adaptation of intrinsic motivation within a lifetime using a framework with two learners. The adaptive intrinsic motivation is a signal that allows the reinforcement-learning mechanism to yield safe exploration policies that lead to efficient learning. This framework suggests that it is possible to view biological agents as lifelong reinforcement learners whose intrinsic motivation depends on their development but who combine that with within-lifetime learning of extrinsic rewards. Ultimately, reinforcement learning addresses the same problem biological agents need to solve, namely learning how to act in an environment in which actions can have better or worse consequences. There therefore is good reason to think that reinforcement learning can contribute to the explanation of lifelong biological learning and behavior.

References

[1] Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, June 2009.

[2] Robert A. Rescorla. Variation in the effectiveness of reinforcement and nonreinforcement following prior inhibitory conditioning. *Learning and Motivation*, 2(2):113–123, May 1971.

[3] Wolfram Schultz. Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80(1):1–27, July 1998.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book - MIT Press, Cambridge, MA, 2018.

[5] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.

[6] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying Count-Based Exploration and Intrinsic Motivation. *arXiv*, November 2016.

[7] Alireza Modirshanechi, He A. Xu, Wei-Hsiang Lin, Michael H. Herzog, and Wulfram Gerstner. The curse of optimism: A persistent distraction by novelty, July 2022.

[8] He A. Xu, Alireza Modirshanechi, Marco P. Lehmann, Wulfram Gerstner, and Michael H. Herzog. Novelty is not Surprise: Human exploratory and adaptive behavior in sequential decision-making. *bioRxiv*, page 2020.09.24.311084, January 2021.

[9] Jack Clark and Dario Amodei. Faulty Reward Functions in the Wild. https://openai.com/blog/faulty-reward-functions/, December 2016.

[10] Keno Juechems and Christopher Summerfield. Where Does Value Come From? *Trends in Cognitive Sciences*, 23(10):836–850, October 2019.

[11] Kate Nussenbaum and Catherine A. Hartley. Understanding the development of reward learning through the lens of meta-learning. *Nature Reviews Psychology*, pages 1–15, April 2024.

[12] Willem E. Frankenhuis and Alison Gopnik. Early adversity and the development of explore–exploit tradeoffs. *Trends in Cognitive Sciences*, 27(7):616–630, July 2023.

[13] Satinder Singh, Richard L Lewis, and Andrew G Barto. Where Do Rewards Come From? page 6, 2009.

[14] Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On Learning Intrinsic Rewards for Policy Gradient Methods. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Annie S. Chen, Archit Sharma, Sergey Levine, and Chelsea Finn. You Only Live Once: Single-Life Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:14784–14797, 2022.

[16] Martin L. Puterman. Chapter 8 Markov decision processes. In *Handbooks in Operations Research and Management Science*, volume 2 of *Stochastic Models*, pages 331–434. Elsevier, January 1990.

[17] Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, and Manuel Kroiss. What Can Learned Intrinsic Rewards Capture? In *International Conference on Machine Learning*, pages 11436–11446. PMLR, 2020.

[18] Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn, 2016.

[19] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv:1611.02779 [cs, stat]*, November 2016.

[20] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, May 1992.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[22] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[23] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256, May 2002.