Can reinforcement learning model learning across development? Online lifelong learning through adaptive intrinsic motivation

Kai Sandbrink (kai.sandbrink@lmh.ox.ac.uk)
University of Oxford, Oxford
United Kingdom

Brian Christian (brian.christian@psy.ox.ac.uk)
University of Oxford, Oxford
United Kingdom

Linas Nasvytis (linasnasvytis@fas.harvard.edu)
Harvard University, Cambridge
United States

Christian Schroeder de Witt (cs@robots.ox.ac.uk)
University of Oxford, Oxford
United Kingdom

Patrick Butlin (patrick.butlin@philosophy.ox.ac.uk)
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.

Keywords: Reinforcement learning, lifelong learning, intrinsic motivation, meta-learning

Introduction

Reinforcement learning (RL), in which an agent learns to optimize expected rewards by interacting with an environment, is a powerful model of biological learning (Niv, 2009). In cognitive science, it traces its origins to theories of operant conditioning and associative learning (Rescorla, 1971), and has predicted neural correlates of learning in neuroscience (Schultz, 1998). In recent decades, it has become a dominant paradigm in machine learning (Sutton & Barto, 2018), achieving milestones such as reaching superhuman performance in Go (Silver et al., 2016), and improving the training of large language models by providing a way to directly incorporate human feedback (Ouyang et al., 2022).

However, reinforcement learning struggles with the so-called *sparse rewards problem*, in which the signal provided by the environment is insufficient to drive learning of complex actions. This is a problem both in machine learning and in the modeling of biological learning. One approach to solving this problem is to introduce intrinsic motivation factors such as count-based intrinsic motivation in machine learning (Bellemare et al., 2016) or novelty- and stochasticity-seeking behavior in humans (Modirshanechi, Xu, Lin, Herzog, & Gerstner, 2022; Xu, Modirshanechi, Lehmann, Gerstner, & Herzog, 2021). These proposed intrinsic motivation signals are usually hand-crafted heuristics that can drive exploratory behavior, skill- and competency-building, or even maintain homeostasis (Oudeyer & Kaplan, 2009).

Hand-crafting intrinsic motivation and intrinsic rewards in machine learning, however, can lead to unpredictable agent behavior (Clark & Amodei, 2016). Similarly, hand-crafting intrinsic motivation features in cognitive science requires numerous assumptions on the part of the experimenter, and will be limited in scope to specific sources and formulations of intrinsic motivation, meaning that it will neglect factors and interactions driving behavior, in particular over longer timescales and open-ended tasks that cannot be as closely controlled.

Biological agents have no acess to hand-crafted intrinsic motivation and reward functions, and must construct their own sense of what is rewarding (Juechems & Summerfield, 2019). Recently, there has been great interest in viewing this process as occurring through meta-reinforcement learning across timescales (Nussenbaum & Hartley, 2024). In this paper, we introduce deep RL networks that address this task, but focus specifically on meta-learning intrinsic motivation while continuing to use an environmentally-determined reward function. This method allows us to focus specifically

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

on the role of meta-learning in determining our algorithms of exploration and action. In this setting, we recapitulate features of changes in intrinsic motivation over the course of development, including a shifting balance from exploration to exploitation, that allow the algorithm to outperform hand-crafted heuristics. In non-stationary tasks, we further show that this form of intrinsic motivation is adaptive to different statistics in the environment. By allowing intrinsic motivation to change freely over time, this method can potentially simulate changes in human patterns of learning and exploration that are impacted by human experiences during development (Frankenhuis & Gopnik, 2023).

This method is compatible with other approaches to address shortcomings of hand-crafted models, such as meta-learning a time-dependent policy or an intrinsic reward function (Singh, Lewis, & Barto, 2009; Zheng, Oh, & Singh, 2018). In theory, they could be combined into a joint model. However, meta-learning intrinsic motivation has several key advantages over meta-learning an intrinsic reward function or directly meta-learning the policy: 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, by reducing the amount of assumptions and designer choices needed in training learning agents on a task *de novo*, it potentially facilitates the use of neural networks to study the time course of learning.

Methods

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 (Chen, Sharma, Levine, & Finn, 2022).

The agent interacts with a Markov decision process (MDP; (Puterman, 1990) $M_{\text{life}} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, p_0, \gamma)$ 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.

Optimal intrinsic motivation

In analogy to previous work on intrinsic rewards (Singh et al., 2009; Zheng, Oh, Hessel, Xu, & Kroiss, 2020), 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.

We address this problem by meta-learning across lifetimes.

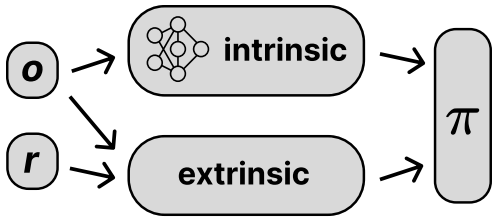

Figure 1: Architecture of the combined intrinsic-extrinsic learning system. Both the intrinsic and the extrinsic systems output a policy on every step that is combined as a weighted average based on intrinsic motivation strength $\alpha \in [0, 1]$. Both the intrinsic and extrinsic learner are trained using rewards, but only the extrinsic learner receives reward information as observations during the episode.

Meta-learning occurs over a set $\mathcal{M}_{\text{evol}}$ of Markov decision processes (MDPs) from which we sample according to a distribution $\rho_{\mathcal{M}_{\text{evol}}} : \mathcal{M} \to \mathbb{R}_+$ at each new lifetime (Wang et al., 2016; Duan et al., 2016). 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.

Concretely, we model three different variations of the ten-armed bandit testbed from Sutton and Barto (2018). In these stateless tasks, the agent has the option of choosing from ten different actions with different payout magnitudes on each step. Episodes are 100 steps long. The first task is the classic stationary ten-armed bandit testbed, in which the payout magnitude of each action is sampled from a standard normal distribution $\mathcal{N}(0, 1)$ at the beginning of each episode. The second is a hardcoded version of the ten-armed bandit testbed, in which the payout magnitude of the first possible action is of a higher payout magnitude $\mathcal{N}10, 1$. The final task is a non-stationary version, in which the payout magnitudes are resampled from $\mathcal{N}0, 1$ during the episode. A parameter called volatility, which is fixed within each episode but varies between them, gives the probability of resampling after every timestep. We train on volatility levels in the intervals $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$, and evaluate on levels across the entire range. In this formulation, $\mathcal{M}_{\text{evol}}$ are all problems in the above sets, and $\mathcal{M}_{\text{life}}$ 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.

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.

Learning extrinsic rewards 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 (Watkins & Dayan, 1992) 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 $\eta = 0.1$.

Learning adaptive intrinsic motivation The *intrinsic learner* has the objective given by Equation 1. 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. (2016). 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; Hochreiter & Schmidhuber, 1997) layer of 64 units followed by a softmax output layer for action selection. We use the REINFORCE algorithm (Williams, 1992) 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.

Combining the two systems Both learners output policies for every timestep. These are combined into one global policy based on a mixture weight $\alpha \in [0,1]$, such that $\pi_{\text{agent}} = (1-\alpha) \times \pi_{\text{extr}} + \alpha \times \pi_{\text{intr}}$ (see Figure 1).

We intentionally set a high value of $\alpha = 0.5$ in order to better study the impact of the intrinsic-motivation system on the agent's overall performance.

Results

Learned intrinsic motivation in the ten-armed bandit testbed

First, we model performance on the standard ten-armed bandit testbed from Sutton and Barto (2018) over episodes of 100 steps. We train the meta-learner over 500,000 episodes. Across five model instantiations, after training, the model reaches an average performance of 87.7 $\pm$ 25.9 (mean $\pm$ SEM over models, see Figure 2A).

We first 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 $\pm$ 3.7 (mean $\pm$ SEM over 100 test episodes) on the same system. Upper Confidence Bound (UCB; Auer, Cesa-Bianchi, & Fischer, 2002), again matching the same weighting as we have between extrinsic and intrinsic motivation, has a performance of 72.9 $\pm$ 3.8 (mean $\pm$ 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).

Evolutionarily-transmitted knowledge

Second, we consider situations where the distribution of bandit payout rates remains constant between different episodes. In this case, the model achieves an average performance of 999.9 $\pm$ 0.36 (mean $\pm$ SEM over models). Figure 2B 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 $\pm$ SEM over 100 test episodes) in this case. The UCB system achieves average rewards of 908.2 $\pm$ 3.2 (mean $\pm$ 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: $t(4) = 229.1$, $p = 1.1e-9$).

Within-lifetime adaptation of exploratory policies

Finally, we show that when given access to extrinsic information (i.e., the rewards that were obtained instead of just action history), the intrinsic motivation supplied by the intrinsic learner adapts accordingly. Figure 2C 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 between different trials. This adaptation encourages greater exploration at levels with higher volatility. At these three volatility settings, across five model instantiations, our network achieves rewards of 47.0 $\pm$ 0.5 (mean $\pm$ SEM over models), 32.0 $\pm$ 0.5 (mean $\pm$ SEM over models), and 12.0 $\pm$ 0.2 (mean $\pm$ SEM over models), respectively. While performance declines with increased volatility, the decrease is not as pronounced as for UCB, which records performance drops to 43.0 $\pm$ 5.0, 33.0 $\pm$ 4.0, and 7.8 $\pm$ 4.0 (mean $\pm$ SEM over 100 test episodes) for the respective settings. The ε-greedy system exhibits even less adaptability to volatility changes, with rewards of 20.0 $\pm$ 5.0, 18.0 $\pm$ 4.0, and 17.0 $\pm$ 5.0 (mean $\pm$ SEM over 100 test episodes) for the three volatility levels. Statistical analyses reveal significant differences in per-

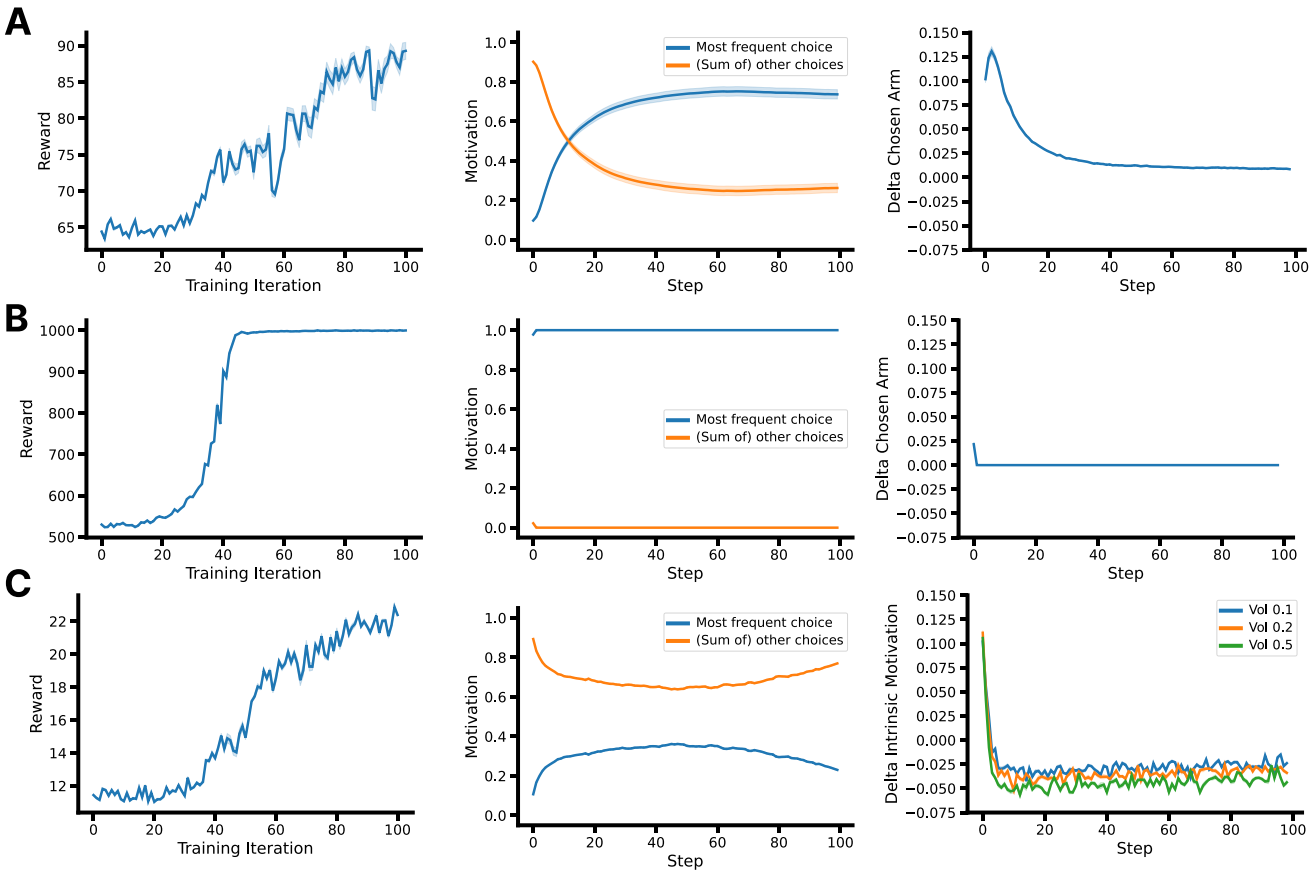

Figure 2: Behaviour of the intrinsic-motivation learning module on variants of the standard 10-armed bandit testbed. **A.** Behavior on the stationary 10-armed bandit task. (*left*) Learning curves as measured on the training distribution for each episode over 1000 instantiations for tasks 1 and 3 and 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.

formance 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.

Comparing learned intrinsic motivation function with hand-crafted heuristics The meta-learned intrinsic motivation follows a smooth transition from encouraging exploration to exploitation: After an exploratory period where the intrinsic motivation is spread across the ten arms, the system switches to habit-driven learning and stabilizes into having chosen one particular arm. In contrast, hand-crafted heuristics will continue to favor exploration even when it is no longer beneficial (Figure 3).

Discussion

RL remains in active use in neuroscience and psychology (Hattori et al., 2023), although it is primarily used to describe learning in controlled experimental conditions over short time scales (Eckstein, Wilbrecht, & Collins, 2021). Using reinforcement learning to describe learning in the real world and over a longer timespan presents significant challenges. In this paper, we provide a potential solution to one unresolved piece of the puzzle, by describing a way to model the origin and development of intrinsic motivation.

Modelling biological learning and behavior with neural networks

Most learning in RL models takes place across numerous episodes in a single, well-defined MDP. Newer RL algorithms have been trained to meta-learn across whole distributions of MDPs "in weights" (Wang et al., 2016; Duan et al., 2016)

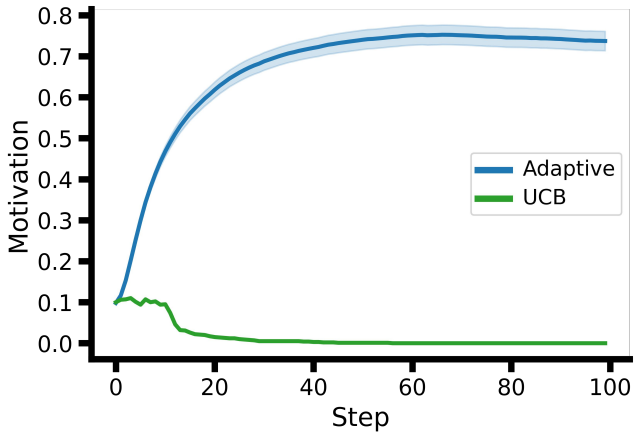

Figure 3: Comparison of (*blue*) meta-learned intrinsic motivation terms with (*green*) hand-crafted heuristics on the stationary ten-armed bandit testbed showing intrinsic motivation attributed to the dominant arm over the course of the whole experiment

and thus have the ability to adapt rapidly to each individual newly-introduced MDP "in context". However, in both cases, learning the original algorithm operates over long timescales and requires much trial and error. These methods are therefore ill-suited to represent learning over the course of a single life, in which episodic resets are impossible. In a single-life setting, there is a significantly greater tension between the imperatives to explore and to avoid excessive risks. In the formulation for robotics used by Chen et al. (2022), the agent has access to prior data $\mathcal{D}$ consisting of transitions from a source MDP $\mathcal{M}_{\text{source}}$ which could come, for instance, from expert trajectories.

In contrast, we model learning as occurring across two different timescales, mirroring the fact that animals both learn within single lifetimes and benefit from evolutionary adaptation. In our setting, the agent has no such explicit knowledge, but can instead call upon intrinsic motivation, which benefits from evolutionary history but may adapt within a single life. The meta-learned intrinsic motivation needs to guide the extrinsic learner to learn a representation of the value function in as safe a way as possible, without having access to environmental rewards itself. In the volatile environments in task 3, the observation in these simulations does include reward information, yet this could be replaced by an environmental change signal, and the task of aiding the extrinsic learner to build a value function of the environment *de novo* in each episode is the same. This task is more difficult than learning to solve the task directly, and, as reported for learned intrinsic rewards by Zheng et al. (2020), we would expect a meta-RL agent trained on the same tasks to outperform our combined agent. However, combined agents allow us to study the learning of agents that begin with a naive value function.

In our setting, we maintain the mixing coefficient α between the extrinsic and intrinsic policy steady at 0.5 over the whole episode, yet this could be set to decrease over time, or also be meta-learned, giving an additional degree of freedom to the intrinsic motivation system. This process of meta-learning hyperparameters could be used even when the extrinsic learner is itself a neural network or other function approximation algorithm, and be extended to include other kinds of learning parameters such as the degree of entropy regularization, learning rate, and similar. This could provide a principled way to study the learning of neural networks on different tasks, without needing to make too many explicit assumptions about the hyperparameters of the inner learning algorithm.

Intrinsic motivation and reward

We adopt a precise definition of 'intrinsic motivation': intrinsic motivation is motivation that does not depend on past rewards received by the agent, although it may depend on past actions. Over longer timescales, however, it is nonetheless ultimately determined by evolutionary pressures. This definition captures the traditional idea that intrinsic motivation is motivation that is not derived from rewards (Ryan & Deci, 2000). As an instance of this idea, we might say that an animal repeatedly performing an action for which it has not been rewarded is evidence of intrinsic motivation. This definition is also consistent with the way that 'intrinsic motivation' is used in some reinforcement learning research. For example, the UCB algorithm for balancing exploitation with exploration in bandit problems includes terms for extrinsic motivation, which does depend on past reward, and intrinsic motivation, which depends only on the number of times an action has previously been performed (Sutton & Barto, 2018). Given this definition, intrinsic motivation must have its source in adaptation across generations, because this is the only way in which it can be sensitive to the adaptive value of forms of behaviour.

In its broadest form, this definition is compatible even with definitions of intrinsic motivation as motivation for open-ended learning that does not have an explicit goal, such as novelty search (Stanley & Lehman, 2015; Lehman & Stanley, 2011). For instance, a system could learn to associate taking novel actions with eventually reaching higher reward during meta-learning, which would then appear like open-ended curiosity from the perspective of the single-life extrinsic agent. However, to model truly generalizable forms of these systems, since the meta-learning system is ultimately also a function approximation, the meta-learning would need to occur over very broad distributions of tasks. Otherwise, the meta-learning system will be liable to the same problems of generalization inherent to neural networks. Given a sufficiently varied training distribution, it would be very interesting to study the kinds of representations that emerge in the intrinsic motivation system. An alternative approach is to constrain the system to particular kinds of representations for specific use cases, for instance using kernel methods for novelty representations (Becker, Modirshanechi, & Gerstner, 2024).

In this work, we have been focusing specifically on intrin-

sic motivation, while referring closely to prior work on intrinsic rewards. The distinction is that intrinsic motivation such as a boost given to less-frequently-explored actions is supplied at the moment of action selection, while intrinsic rewards are typically supplied after an action has been taken. In a model-based setting, this distinction partially collapses, since the agent can use planning to simulate how much intrinsic rewards an action is expected to give before taking it. In a model-free setting, however, the implementation makes a difference, since only motivation that has an impact before an action is taken can stop an agent from taking dangerous actions and guide safe exploration. This is particularly relevant in a single-life setting. Since human learning is thought to have both model-free and model-based components (Gershman, 2017), it makes sense to also consider the motivation that is useful in this setting.

Another difference is that a learning agent supplied with intrinsic rewards, in addition to extrinsic rewards, will learn a single value function or policy that jointly addresses both sources of rewards. Having a separate system allows for an interpretable representation, that could be more flexibly modified in response to environmental changes.

Outstanding conceptual challenges

Our work helps to address one challenge to RL as a model of biological agency, which is that models must capture the combination of within-lifetime learning and evolutionary adaptation. However, this approach still faces considerable conceptual challenges, some of which are related to our work.

One challenge concerns the relationship between reinforcement learning agents and biological agents. Barto and colleagues have argued that, because biological agents do not receive a reward signal from the environment, they do not face the problem studied in reinforcement learning research, and are therefore not correctly thought of as reinforcement learning agents. Their view is that reinforcement learning agents exist as homunculi within animal minds, which work to maximise whatever reward signals they are given by other parts of the mind (Barto, Singh, & Chentanez, 2004; Singh et al., 2009; Barto, 2013). This conception is arguably supported by models like ours in which the extrinsic learner, which engages in reinforcement learning, is only one part of the agent. The agent would still require exploration and other forms of intrinsic motivation in order to drive its learning. However, the advantages and disadvantages of this conception have yet to be fully explored. It is also crucial to appreciate that value for biological agents is not wholly subjective, and that animals themselves do face the problem of learning from experience to achieve valuable outcomes.

A second challenge is to make sense of reward functions in biological agents. One issue is that if rewards are constructed then they are more difficult to disentangle from the agent's representation of the value function or policy; this relates to one possible definition of the reward function, which is as a description of a signal that acts as an input to reinforcement learning. Reward functions can also be thought of as optimisation targets, but if behaviour is the product of multiple sources of motivation, there may be no one optimisation target for behaviour (a related but distinct question is whether the values of options are represented in a common currency; Levy & Glimcher, 2012; Spurrett, 2016).

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.

Acknowledgments

We would like to thank the Principles of Intelligent Behavior in Biological and Social Systems (PIBBSS) summer research fellowship, where this project was conceived and initial simulations conducted. K.J.S. was additionally funded by a Cusansuwerk Doctoral Research Fellowship.

References

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002, May). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, *47*(2), 235–256. doi: 10.1023/A:1013689704352

Barto, A. G. (2013). Intrinsic Motivation and Reinforcement Learning. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems* (pp. 17–47). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-32375-1$_2$

Barto, A. G., Singh, S., & Chentanez, N. (2004). Intrinsically Motivated Learning of Hierarchical Collections of Skills.

Becker, S., Modirshanechi, A., & Gerstner, W. (2024). Representational similarity modulates neural and behavioral signatures of novelty. *bioRxiv*, 2024–05.

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016, November). Unifying Count-Based Exploration and Intrinsic Motivation. *arXiv*.

Chen, A. S., Sharma, A., Levine, S., & Finn, C. (2022). You Only Live Once: Single-Life Reinforcement Learning. *Advances in Neural Information Processing Systems*, *35*, 14784–14797.

Clark, J., & Amodei, D. (2016, December). *Faulty Reward Functions in the Wild.* https://openai.com/blog/faulty-reward-functions/.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016, November). RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv:1611.02779 [cs, stat]*.

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021, October). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137. doi: 10.1016/j.cobeha.2021.06.004

Frankenhuis, W. E., & Gopnik, A. (2023, July). Early adversity and the development of explore–exploit tradeoffs. *Trends in Cognitive Sciences*, *27*(7), 616–630. doi: 10.1016/j.tics.2023.04.001

Gershman, S. J. (2017). *Reinforcement Learning and Causal Models* (Vol. 1; M. R. Waldmann, Ed.). Oxford University Press. doi: 10.1093/oxfordhb/9780199399550.013.20

Hattori, R., Hedrick, N. G., Jain, A., Chen, S., You, H., Hattori, M., … Komiyama, T. (2023, November). Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience*, 1–10. doi: 10.1038/s41593-023-01485-3

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Juechems, K., & Summerfield, C. (2019, October). Where Does Value Come From? *Trends in Cognitive Sciences*, *23*(10), 836–850. doi: 10.1016/j.tics.2019.07.012

Lehman, J., & Stanley, K. O. (2011). Novelty Search and the Problem with Objectives. In R. Riolo, E. Vladislavleva, & J. H. Moore (Eds.), *Genetic Programming Theory and Practice IX* (pp. 37–56). New York, NY: Springer. doi: 10.1007/978-1-4614-1770-5_3

Levy, D. J., & Glimcher, P. W. (2012, December). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. doi: 10.1016/j.conb.2012.06.001

Modirshanechi, A., Xu, H. A., Lin, W.-H., Herzog, M. H., & Gerstner, W. (2022, July). *The curse of optimism: A persistent distraction by novelty.* bioRxiv. doi: 10.1101/2022.07.05.498835

Niv, Y. (2009, June). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. doi: 10.1016/j.jmp.2008.12.005

Nussenbaum, K., & Hartley, C. A. (2024, April). Understanding the development of reward learning through the lens of meta-learning. *Nature Reviews Psychology*, 1–15. doi: 10.1038/s44159-024-00304-1

Oudeyer, P.-Y., & Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *1*, 6. doi: 10.3389/neuro.12.006.2007

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022). *Training language models to follow instructions with human feedback* (Vol. 35).

Puterman, M. L. (1990, January). Chapter 8 Markov decision processes. In *Handbooks in Operations Research and Management Science* (Vol. 2, pp. 331–434). Elsevier. doi: 10.1016/S0927-0507(05)80172-0

Rescorla, R. A. (1971, May). Variation in the effectiveness of reinforcement and nonreinforcement following prior inhibitory conditioning. *Learning and Motivation*, *2*(2), 113–123. doi: 10.1016/0023-9690(71)90002-6

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. doi: 10.1037/0003-066X.55.1.68

Schultz, W. (1998, July). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, *80*(1), 1–27. doi: 10.1152/jn.1998.80.1.1

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016, January). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi: 10.1038/nature16961

Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where Do Rewards Come From? , 6.

Spurrett, D. (2016, November). Common Currencies, Multiple Systems and Risk Cognition: Evolutionary Trade-offs and the Problem of Efficient Choices. *Journal of Cognition and Culture*, *16*(5), 436–457. doi: 10.1163/15685373-12342187

Stanley, K. O. O., & Lehman, J. (2015). *Why Greatness Cannot Be Planned: The Myth of the Objective* (2015th edition ed.). Cham Heidelberg New York Dordrecht London: Springer.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: A Bradford Book - MIT Press.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., … Botvinick, M. (2016). *Learning to reinforcement learn* (No. arXiv:1611.05763). arXiv.

Watkins, C. J. C. H., & Dayan, P. (1992, May). Q-learning. *Machine Learning*, *8*(3), 279–292. doi: 10.1007/BF00992698

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*, 229–256.

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., & Herzog, M. H. (2021, January). Novelty is not Surprise: Human exploratory and adaptive behavior in sequential decision-making. *bioRxiv*, 2020.09.24.311084. doi: 10.1101/2020.09.24.311084

Zheng, Z., Oh, J., Hessel, M., Xu, Z., & Kroiss, M. (2020). What Can Learned Intrinsic Rewards Capture? In *International conference on machine learning* (pp. 11436–11446).

Zheng, Z., Oh, J., & Singh, S. (2018). On Learning Intrinsic Rewards for Policy Gradient Methods. *Advances in Neural Information Processing Systems*, *31*.