Exploitation Is All You Need... for Exploration

Micah Rentschler
Vanderbilt University
Nashville, TN, USA
micah.d.rentschler@vanderbilt.edu

Jesse Roberts
Tennessee Technological University
Cookeville, TN, USA
jtroberts@tntech.edu

***Abstract*—Exploration is a central challenge when learning in novel situations. Conventional solutions to the exploration–exploitation dilemma inject explicit incentives such as randomization, uncertainty bonuses, or intrinsic rewards to encourage exploration. In this work, we hypothesize that an agent trained with reinforcement learning (RL) to maximize a purely greedy exploitation objective can exhibit exploratory behavior, provided three conditions are met: (1) Recurring Environmental Structure, where the environment features repeatable regularities that allow past experience to inform future choices; (2) Agent Memory, enabling the agent to retain and use historical interaction data; and (3) Long-Horizon Credit Assignment, where learning propagates returns over a time frame sufficient for the delayed benefits of exploration to impact current decisions. Through experiments in stochastic multi-armed bandits and temporally extended gridworlds, we verify that, when both structured tasks and agent memory are present, a policy trained on a strictly greedy objective exhibits information-seeking exploratory behavior. These findings suggest that, under the right prerequisites, exploration and exploitation need not be treated as orthogonal objectives but can emerge from a unified reward-maximization process.**

***Index Terms*—exploration–exploitation, meta-reinforcement learning, transformers, emergent exploration.**

I. Introduction

Consider a sailor marooned on an unfamiliar island. To survive, he must quickly find reliable sources of fresh water and food. Early on, he may spend precious energy scouting inland, checking different coves, and testing plants—not because wandering is rewarding in itself, but because the information he gathers can make future decisions more profitable and less risky. Over time, as he learns which paths and resources are dependable, his behavior shifts from broad search to efficient routines that maximize the utility of his acquired knowledge. Can machines do the same? Can they *learn* to explore in order to exploit?

Exploration is ubiquitous in humans and animals, yet unlike drives such as hunger or pain, there is no clear, biologically plausible mechanism directly incentivizing it. Some neuroscientific work suggests that the brain balances exploration and exploitation through meta-learning, which relies on repeating tasks, memory, and rewards [1].

In reinforcement learning (RL), designers face a similar exploration–exploitation dilemma: how should an agent balance exploring unknown actions (potentially yielding future gain) and exploiting known rewards [2], [3]? Traditional RL methods address this by introducing explicit exploratory incentives such as ϵ-greedy sampling [2], Upper Confidence Bound (UCB) algorithms [4], or intrinsic-motivation mechanisms like curiosity-driven exploration [5], [6] that treat exploration as orthogonal to exploitation.

In contrast, we ask: *Can exploration emerge organically from a purely greedy objective?* We hypothesize that it can under three key conditions:

1) **Recurring Environmental Structure.** The task repeats, so early-acquired information remains valuable later.
2) **Agent Memory.** The policy retains and utilizes past experiences across episodes.
3) **Long-Horizon Credit Assignment.** Learning must connect information gathering to long-term payoffs.

We empirically test this hypothesis by training a meta-RL agent in both multi-armed bandits and temporally extended gridworlds. With all three conditions met, our results show that the greedy agent consistently exhibits information-seeking behavior even without explicit incentives. Removing structure or memory eliminates exploration. Long credit horizons, on the other hand, appear necessary only in certain settings.

Our findings challenge the view that exploration and exploitation require separate objectives. In repetitive environments with agent memory and long-term credit assignment, reward maximization alone enables effective exploration. Instead of engineered exploration incentives, we believe it is more effective to train with the ingredients that motivate exploration in the first place.

Training code will be made publicly available upon acceptance of this paper.

II. Related Work

The exploration–exploitation trade-off is a longstanding challenge in reinforcement learning (RL) and sequential decision making [2], [3], [7]. Early approaches focused on randomization techniques, such as ϵ-greedy or softmax action selection, as well as principled methods like Upper Confidence Bound (UCB) [4] and Thompson Sampling (TS) [8], which provide theoretical regret guarantees.

Instead of explicitly biasing the action distribution towards exploration, intrinsic motivation methods, such as curiosity-driven exploration, add reward for visiting novel or unpredictable states [5], [9]. Random Network Distillation (RND) [10] is a notable example, providing intrinsic bonuses in sparse-reward Atari games. Recent methods, such as SPIE [11], combine counts and trajectory structure to target bottlenecks in exploration. VariBAD [12] applies Bayesian principles to

guide implicit exploration, though with explicit modeling of uncertainty.

In the meta-RL literature, recurrent policies have demonstrated what appears to be exploratory behavior [13], [14]. However, the conditions under which this behavior emerges have not, to our knowledge, been systematically studied. This motivates us to examine whether meta-RL can resolve the exploration–exploitation trade-off and to ask why, and under what conditions, it does so.

Our inquiry is further motivated by neuroscientific evidence suggesting that the brain's meta-learning mechanisms manage the exploration–exploitation trade-off by relying on specific prerequisites: recurring task structures, memory, and reward-driven learning processes [1]. Notably, these prerequisites correspond closely to the conditions we hypothesize are essential for emergent exploration.

In this work, we build on the observation that meta-RL agents can exhibit exploratory behavior [13]–[15] and ask why this behavior emerges and under what conditions. We systematically delineate and validate the conditions under which exploration arises, emphasizing the critical roles of recurring environmental structure and agent memory. Our analysis also raises new questions about long-horizon credit assignment, suggesting that its importance may depend on the setting. These insights may prove useful as we move toward a future in which RL agents can decide for themselves what to explore.

III. Background

A. Repeated Markov Decision Processes (MDPs)

To study when exploration can emerge from pure exploitation, we adopt a repeated-task setting. A fixed, partially observable Markov decision process (POMDP)

$$M = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, \Omega, r, \gamma)$$

is sampled from a distribution, and the agent interacts with this *same* environment for multiple episodes. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, and $\mathcal{O}$ the set of observations available to the agent. The transition dynamics are given by P, the observation kernel by Ω, the reward function by r, and γ is the discount factor.

At the end of each episode, with probability $1/n$, a new parameterization of the environment is sampled, marking the start of a new task block. Thus, each block consists of a geometrically distributed number of episodes with mean n, during which the environment's parameters remain fixed. As a result, information gathered early in the block—such as the location of a goal or the identity of the best arm—remains valuable in later episodes. This regime enables agents to accumulate and exploit knowledge across episodes.

B. Meta-Reinforcement Learning (Meta-RL)

Traditional reinforcement learning (RL) trains agents to solve a single environment. In contrast, meta-reinforcement learning (Meta-RL) aims to produce agents that can rapidly adapt to new tasks. One approach, Model-Agnostic Meta-Learning (MAML), seeks an initialization that enables efficient weight updates at test time [16]. By contrast, RL^2-style methods avoid weight updates at test time. Instead, they employ recurrent neural networks (or transformers) to process the history of interactions and infer strategies for previously unseen environments, effectively "learning to learn" [13], [14]. In this paper, we use the term meta-RL to refer to this second variant.

IV. Methodology

Our experiments are designed to empirically test the hypothesis that exploration can emerge from pure exploitation objectives under certain preconditions. We use both short-horizon (bandit) and long-horizon (gridworld) environments with a repeated-MDP structure, and we train a transformer-based meta-RL agent offline using data collected from a random policy. Training is performed using the standard Deep Q-Network (DQN) algorithm [17]. Although DQN is not typically associated with meta-learning, combining it with a model conditioned on historical context yields a meta-RL agent [15], [18].

A. Environments

We evaluate our hypothesis in two classes of environments: stochastic multi-armed bandits and temporally extended gridworlds. Both are framed within the repeated-task regime described in Section III, where environment parameters remain fixed across multiple episodes within a block.

- **Multi-Armed Bandits:** Bandit environments are single-step decision problems, classically used to study the

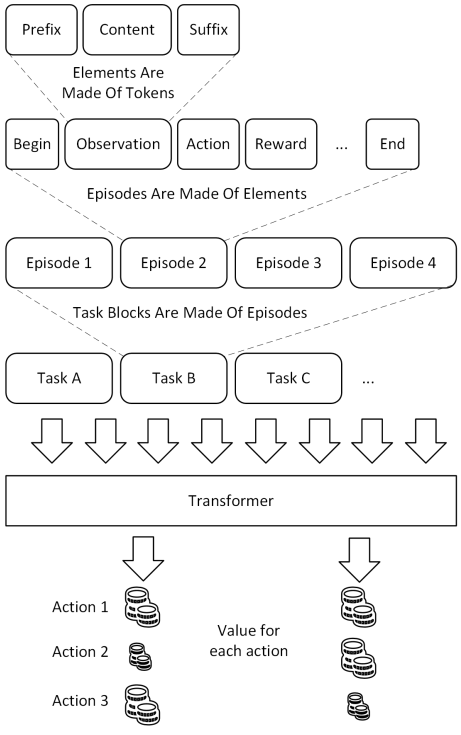


Fig. 1. Setup for DQN meta-RL learning with a transformer. Tokens make up elements, which comprise episodes, which in turn make up task blocks. This is fed to the transformer which is trained to output the value of each action.

exploration–exploitation trade-off. We adopt the K-armed bandit setting, where each arm provides rewards from a stationary distribution (Bernoulli in our experiments). Each episode consists of a single step. Reward parameters remain fixed within a block. After each episode, a new block begins with probability $1/n$. Thus, block lengths follow a geometric distribution with mean n. Larger n yields greater recurrence—information from earlier episodes remains useful later—while smaller n reduces recurrence. This design allows us to isolate the effect of environmental repetition on emergent exploration.

- **Gridworlds:** To study exploration in multi-step tasks, we use variants of the Frozen Lake environment. The agent begins in a start state and must reach a goal while avoiding holes, receiving reward $1/t$, where t is the number of steps required to succeed. As in the bandit case, task blocks persist for a geometrically distributed number of episodes with mean n, after which a new map is sampled. Each map specifies random start, goal, and hole positions; trivial maps (goal reachable in fewer than three moves) and impossible maps (goal unreachable) are discarded. Grid sizes are sampled uniformly between 3×3 and 5×5, requiring agents to generalize across diverse maps without prior knowledge of structure.

Episodes are represented by stacking the sequence of actions, observations, and rewards. Episodes are then stacked into blocks, and blocks are further stacked until the transformer's context is filled (see Figure 1).

B. Agent Architecture

Agents use transformer-based value functions following [18]. A historical sequence consisting of states, actions, rewards, and termination bits is converted into ASCII text. This text is then tokenized using the Llama tokenizer. In this representation, one time step corresponds to about 9–15 tokens. The model ingests the most recent X tokens, which defines its effective memory capacity: smaller X restricts memory to recent events, whereas larger X enables cross-episode retention within a block. We initialize from the pretrained Llama 3.2 3B model and apply LoRA adapters (r_{LoRA}=32, α_{LoRA}=32) for efficient fine-tuning.

C. Training

We train with DQN using a delayed target network for stability updated via Polyak averaging with factor τ=0.1. The discounting scheme is tailored to the repeated-task (block) setting to control temporal credit assignment across episodes within a block:

- Within an episode, per-step discount is set to 1.
- On the terminal transition of an episode, we apply $\gamma_{\text{episode}} \in [0, 1]$.
- On the terminal transition of a block boundary, we set the discount to 0.

Thus, unlike the usual practice of setting the terminal-episode discount to 0, we allow nonzero cross-episode credit *within* a block when $\gamma_{\text{episode}}>0$.

Unless otherwise noted, training runs for 400 iterations in bandits and 1200 iterations in gridworlds. Training is conducted offline: we generate M data streams by running environments with a random policy until each stream contains at least $20X$ tokens. At each iteration, we sample a fixed number of streams to form a batch, keeping tokens-per-batch approximately constant across context sizes. We apply action masking so that only valid tokens contribute to the DQN loss.

D. Testing

Prior to evaluation, we seed the transformer's context with trajectories generated by a random policy in other environments. At test time, a new environment is sampled, and the model interacts for a fixed number of episodes. No additional exploration noise is introduced.

E. Metrics

Exploration is valuable only insofar as it translates into higher future reward. For this reason, our primary evaluation metric is the cumulative return achieved over a limited number of episodes. Policies that engage in effective exploration discover superior strategies and thereby accumulate greater total reward, whereas policies that fail to explore typically overfit to incidental experiences and converge to suboptimal returns. Measuring return directly therefore provides the most natural and principled way to assess the quality of exploration.

To complement this outcome-based perspective, we also analyze behavioral indicators of exploration. In the gridworld experiments, we visualize state visitation distributions. These heatmaps serve as heuristic evidence of exploration by revealing whether agents initially spread their visits broadly across the map before converging to efficient goal-directed paths. Such qualitative patterns allow us to compare emergent exploration to human-like strategies, highlighting whether agents exhibit the same exploratory-to-exploitative progression commonly observed in human learning.

F. Baselines

Our aim is not to surpass conventional baselines in raw performance, but to show that emergent exploration through greedy exploitation crucially depends on recurrence, memory, and temporal horizon. Still, it is important to situate our results relative to established exploration strategies in order to calibrate their significance.

For the bandit environments, we therefore compare against canonical baselines: Thompson Sampling, ϵ-greedy, a random policy, and an oracle with full knowledge of reward distributions. These comparisons allow us to position emergent exploration alongside both theoretically principled methods (Thompson Sampling) and widely used heuristics (ϵ-greedy), while also bracketing performance between lower (random) and upper (oracle) bounds.

In the gridworld setting, there are no widely accepted meta-RL baselines that directly isolate exploration. A central advantage of meta-RL is its ability to generalize exploratory strategies to multi-step domains. Comparing against standard

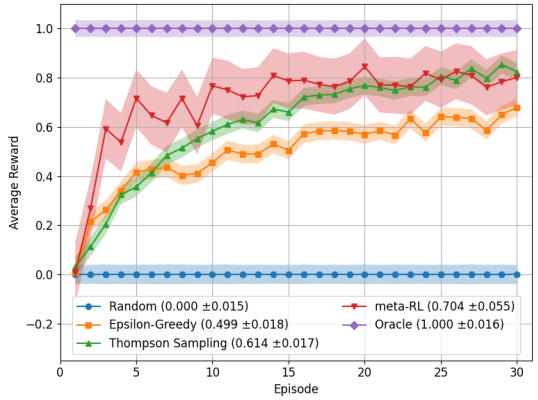


Fig. 2. Reward per episode in the 3-arm bandit environment, comparing meta-RL ($n = 30$ $X = 1024$ $\gamma_{episode} = 0.9$) to baseline strategies. Shaded areas denote 95% confidence intervals. As Meta-RL interacts with the environment, its performance improves, demonstrating an ability to explore and then exploit.

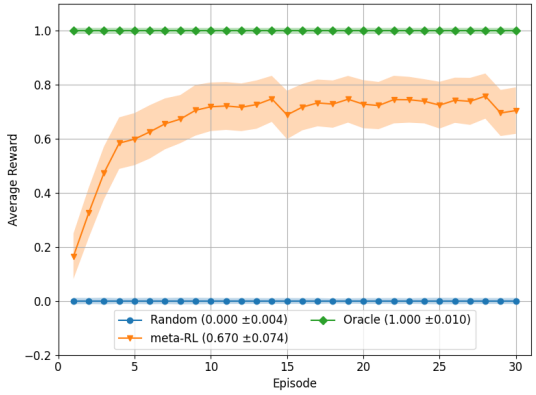


Fig. 3. Reward per episode in the Frozen Lake gridworld, comparing meta-RL ($n = 30$ $X = 1024$ $\gamma_{episode} = 0.9$) to random and oracle strategies. Shaded areas denote 95% confidence intervals. The meta-RL agent achieves significant gains, demonstrating that it gathers information (exploration) that is later used to gather rewards (exploitation).

RL baselines would require online parameter updates during evaluation, introducing confounding factors such as update frequency, stability, and computational overhead—issues that are orthogonal to our study. To avoid these complications, we restrict our comparisons to two clear references: a random policy and an oracle policy.

V. Experiments

We evaluate our hypothesis—that exploration can emerge from pure exploitation given the right structural conditions—using controlled ablation studies in multi-armed bandit and gridworld tasks. Across all experiments, we systematically vary: the degree of environment recurrence (n), agent memory capacity (X), and the temporal credit assignment horizon ($\gamma_{episode}$).

All results are averaged over multiple seeds: 10,000 runs for baselines and 1,000 for meta-RL in the bandit task, and 1,000 runs for baselines and 100 for meta-RL in gridworld. To demonstrate significance, we report 95% confidence intervals taken across all seeds.

To facilitate comparison and interpretation, we linearly normalize all reported rewards: a value of 1 denotes performance of an oracle agent, while 0 indicates the average performance of a random agent under identical conditions.

A. Multi-Armed Bandit Results

Figure 2 shows the performance of the meta-RL agent in a 3-arm bandit environment. In the 30-episode regime, the agent's cumulative returns are on par with the Thompson Sampling baseline.

a) Effect of Recurring Structure: Table I shows that higher mean block lengths (n)—i.e., greater environment recurrence—enable agents to leverage early exploration for improved performance in later episodes. Performance drops sharply for smaller n, confirming that recurring structure is necessary for emergent exploration.

b) Effect of Memory Capacity: Table II demonstrates that reducing the agent's memory capacity (context window X) leads to a sharp decline in cumulative reward. Below a critical threshold, exploration fails to emerge, confirming that sufficient memory is essential.

c) Effect of Temporal Credit Assignment: When the temporal horizon is ablated ($\gamma_{episode} = 0$), meta-RL continues to display exploratory behavior, without appreciable reduction in performance. This result is unexpected and discussed further in Section VI.

B. Gridworld Results

Figure 3 shows the performance of the meta-RL agent in the Frozen Lake gridworld. The agent learns exploratory policies improving from random toward the expert oracle.

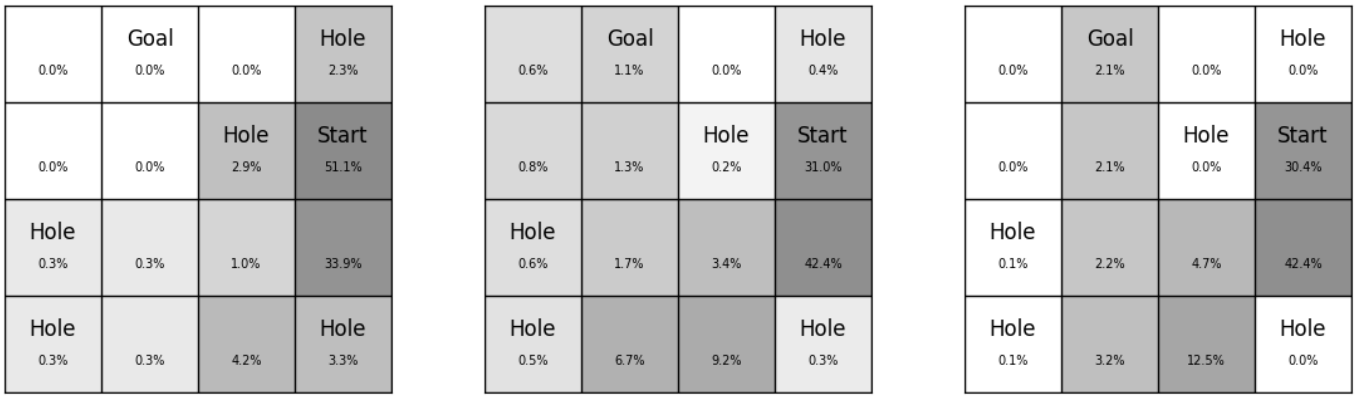
a) Effect of Recurring Structure: As with bandits, increasing block length n in gridworlds yields higher cumulative

TABLE I
CUMULATIVE REWARD AS A FUNCTION OF MEAN BLOCK LENGTH n IN THE 3-ARMED BANDIT ENVIRONMENT.

n	ϵ-Greedy	TS	Meta-RL
30	0.499 ±0.018	0.614 ±0.017	0.704 ±0.055
10	0.354 ±0.020	0.339 ±0.019	0.509 ±0.064
3	0.147 ±0.027	0.092 ±0.025	0.325 ±0.082
1	-0.083 ±0.041	-0.066 ±0.041	0.043 ±0.130

TABLE II
CUMULATIVE REWARD AS A FUNCTION OF CONTEXT WINDOW X IN THE 3-ARMED BANDIT ENVIRONMENT.

X	ϵ-Greedy	TS	Meta-RL
1024			0.704 ±0.055
256			0.792 ±0.054
128	0.499 ±0.018	0.614 ±0.017	0.574 ±0.062
64			0.547 ±0.060
32			-0.052 ±0.099



(a) Early (b) Middle (c) Late

Fig. 4. State visitation heatmaps in the same gridworld (early episodes 1–3, middle 4–7, and late 8–30). Darker shading indicates more time spent in that state. Early episodes show exploration around the start; middle episodes avoid holes and explore further out; late episodes follow discovered paths to goal.

reward (Table III). The agent exploits repeated structure to improve over time.

b) Effect of Memory Capacity: Reducing the context window X also decreases agent performance in gridworlds, though the critical threshold for collapse occurs at higher X than in bandits due to the increased episode length (Table IV).

c) Effect of Temporal Credit Assignment: Our results show that while meta-RL agents can exhibit exploratory behavior in bandit tasks even with $\gamma_{episode} = 0$, in gridworld environments increasing the discount factor from zero to a nonzero value leads to a moderate improvement in overall performance, suggesting reward-induced exploration, with the average reward rising from 0.408 ±0.089 to 0.670 ±0.074.

d) State Visitation: Figure 4 shows the state visitation distribution across block progression, aggregated over 10 trials in the same environment class with different random seeds (percentages represent the relative number of time steps spent in a particular state). Early on, agents visit a broad range of states. As training progresses, visitation becomes concentrated along paths leading to the nearest goal, and agents appear to avoid hazardous states such as holes. Notably, exploration seems to expand outward from the starting position, gradually covering more distant regions.

TABLE III
CUMULATIVE REWARD AS A FUNCTION OF MEAN BLOCK LENGTH n IN FROZEN LAKE GRIDWORLDS.

n	Meta-RL
30	0.670 ±0.074
10	0.441 ±0.070
3	0.254 ±0.071
1	-0.050 ±0.004

TABLE IV
CUMULATIVE REWARD AS A FUNCTION OF MEMORY WINDOW X IN FROZEN LAKE GRIDWORLDS.

X	Meta-RL
1024	0.670 ±0.074
256	0.120 ±0.056
128	0.047 ±0.026
64	-0.001 ±0.033

VI. DISCUSSION

Our findings provide evidence that exploration does not require explicit bonuses or injected randomness. Instead, it can emerge naturally from purely exploitative training when the right conditions are met. Specifically, policies can exhibit information-seeking behavior when environments have recurring patterns, agents possess sufficient memory, and learning propagates reward across a sufficiently long horizon.

In bandit tasks, exploration persists even when cross-episode reward propagation is disabled. We hypothesize that the agent's context contains sufficient stochasticity to make its choices across the K arms diverse enough to explore the space without being driven by cross-episode rewards. This seems to break down in gridworlds, where rewards depend on a sequence of actions: presumably, the model's output distribution is not broad enough to support effective exploration, so cross-episode credit assignment becomes beneficial.

Overall, our experiments suggest that recurring patterns in the environment and sufficient agent memory are necessary, and that long-horizon credit assignment may help elicit exploration in temporally extended settings.

VII. LIMITATIONS

Our study focuses on environments with explicit recurring structure, sufficient agent memory, and long-term credit propagation. While we argue that these conditions are necessary, they may not be sufficient; exploration could still fail to emerge for other reasons. Also, further research is needed to evaluate the robustness and scalability of emergent exploration in more complex and diverse domains.

VIII. Conclusion

We identify and validate three factors that influence the emergence of exploration under greedy training: recurring structure, agent memory, and long-horizon credit assignment. In both bandit environments and gridworlds, we show that removing recurring structure or agent memory dramatically reduces exploration. Long-horizon credit assignment has minimal effect in bandit environments, but a significant effect in gridworlds.

Our findings challenge the view that exploration must be manually incentivized. Instead of engineered exploration incentives, we argue that it is more effective to train with the ingredients that motivate exploration in the first place.

We hope this study encourages continued investigation into how context, memory, and long-horizon credit assignment contribute to emergent exploration.

Acknowledgments

The authors would like to acknowledge that ChatGPT helped check and rephrase this manuscript. While all substantive ideas were generated by the authors, the use of generative AI was an invaluable time saver.

References

[1] M. Botvinick, J. X. Wang, J. Kwisthout, and et al., "Prefrontal cortex as a meta-reinforcement learning system," *Nature Neuroscience*, vol. 21, no. 6, pp. 860–868, 2018.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, 2018.

[3] S. Thrun, "Efficient exploration in reinforcement learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 1992, pp. 433–440.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[5] D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, and J. Malik, "Curiosity-driven exploration by self-supervised prediction," in *International Conference on Learning Representations (ICLR)*, 2019.

[6] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, "Large-scale study of curiosity-driven learning," *arXiv preprint arXiv:1808.04355*, 2018.

[7] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[8] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[9] P.-Y. Oudeyer and F. Kaplan, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[10] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, "Exploration by random network distillation," in *International Conference on Learning Representations (ICLR)*, 2019.

[11] C. Yu, N. Burgess, M. Sahani, and S. J. Gershman, "Successor-predecessor intrinsic exploration," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson, "VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[13] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "RL^2: Fast reinforcement learning via slow reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[14] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick, "Learning to reinforcement learn," in *Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci)*, 2017.

[15] L. C. Melo, "Transformers are meta-reinforcement learners," in *International Conference on Machine Learning (ICML)*, 2022, pp. 15 340–15 359.

[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *CoRR*, vol. abs/1703.03400, 2017. [Online]. Available: http://arxiv.org/abs/1703.03400

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[18] M. Rentschler and J. Roberts, "RL + transformer = a general-purpose problem solver," in *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, E. Kamalloo, N. Gontier, X. H. Lu, N. Dziri, S. Murty, and A. Lacoste, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 401–410. [Online]. Available: https://aclanthology.org/2025.realm-1.29/