

# A Novel Framework for Uncertainty-Driven Adaptive Exploration

Leonidas Bakopoulos  
 Technical University of Crete  
 Chania, Greece  
 lbakopoulos@tuc.gr

Georgios Chalkiadakis  
 Technical University of Crete  
 Chania, Greece  
 gchalkiadakis@tuc.gr

## ABSTRACT

Adaptive exploration methods learn complex policies via alternating between exploration and exploitation. An important question for such methods is to determine the appropriate moment to switch between exploration and exploitation and vice versa. This is critical in domains that require the learning of long and complex sequences of actions. In this work, we present a generic adaptive exploration framework that employs uncertainty to address this important issue in a principled manner. Our framework includes previous adaptive exploration approaches as special cases. Moreover, it can incorporate any uncertainty-measuring mechanism of choice, such as mechanisms used in intrinsic motivation, or epistemic uncertainty-based exploration methods; and is experimentally shown to give rise to adaptive exploration strategies that outperform standard ones across several environments. Moreover, we showcase its potential for utilization in safety-critical domains. The code for this work can be found at <https://github.com/leoBakop/ADEU>

## KEYWORDS

Deep Reinforcement Learning; Adaptive Exploration; Uncertainty-Driven Exploration

### ACM Reference Format:

Leonidas Bakopoulos and Georgios Chalkiadakis. 2026. A Novel Framework for Uncertainty-Driven Adaptive Exploration. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 9 pages. <https://doi.org/10.65109/HQII4250>

## 1 INTRODUCTION

Over the past decade, *deep reinforcement learning (DRL)* [19, 23, 37] has become the policy learning method of choice in a variety of domains. In most cases, the goal of online DRL is to approximate the  $Q$ -function via alternating between generating new experiences by interacting with the environment (exploration) and updating the strategy using this experience (learning) [27].

Now, robotics is considered a standard domain for testing DRL [12, 18, 19, 22]. In general, such domains are characterized by: (i) continuous and multi-dimensional *action* space; (ii) informative, multi-component, and *dense* reward function (iii) episodes that are very probable to early termination (or “truncation”) when an agent is set in an unsafe state; (iv) transitions being highly deterministic. In

such domains, the main aim is to learn a complex and long trajectory, necessitating the execution of a sequence of actions that will lead the agent to a specific state.

In domains with the aforementioned characteristics, it is critical for an agent to pose the question: “Is it *always* better to search for new states, or should I carefully choose *when* to explore?” To further understand the importance of correctly answering that question, let us consider the classic Walker domain [36]. In this domain, a two-dimensional bipedal robot aims to walk in the forward direction. At the beginning of an episode, the agent is set to an initial state. An episode is truncated when the agent finds itself in an unsafe state. Suppose also that the agent has already learned an incomplete and close-to-optimal trajectory which, in this case, corresponds to performing a single step. This trajectory can guide the agent from the initial state to a new safe one. The agent’s goal is to extend this trajectory as much as possible. For that to happen, two phases are required: (i) the exploitation of the current policy by replicating the trajectory until its end (by replicating in other words the “single step” trajectory); and (ii) the exploration from the end of the trajectory until finding the next best action (explore after performing a “single step”). We should note that the switch between the two phases should happen within each episode. In case the agent explores instead of replicating the known trajectory, it is more probable to transition to a worse—and maybe unsafe—state rather than the one included in the trajectory. Given this example, Pislár et al. [27] argue that it is more advantageous to ask “*when* to explore” rather than “*how* to explore”.<sup>1</sup>

In this paper, we present *ADEU - ADaptive Exploration via Uncertainty*, a generic exploration framework that decides in a *principled* manner *when* the switching between exploration and exploitation and vice versa should take place within an episode. Intuitively, ADEU enables exploration when it adjudges that its *uncertainty* regarding its policy is “high enough”. Now, uncertainty is arguably a subjective term and is highly dependent on the formulation of the problem being solved. ADEU takes that into account. It can, in fact, incorporate *any* mechanism for measuring uncertainty in its action selection mechanism. Intuitively, such an *uncertainty-measuring* mechanism corresponds to a particular “uncertainty definition”. Moreover, ADEU does not resort to the use of heuristic techniques or thresholds to answer the aforementioned “*is uncertainty high enough*” question: instead, it selects an action in a principled manner, by sampling from a distribution whose spread is proportional to its uncertainty regarding the policy. By being able to incorporate any uncertainty-measuring mechanism, and by selecting an action in the aforementioned principled manner, ADEU is, in fact,



This work is licensed under a Creative Commons Attribution International 4.0 License.

*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems ([www.ifaamas.org](http://www.ifaamas.org)). <https://doi.org/10.65109/HQII4250>

<sup>1</sup>The latter is in fact the question that *deep exploration* methods [5, 20, 21, 25, 29] focus on. Even though these were designed for use in structurally different domains (e.g., video game domains with sparse, discrete rewards, and stochastic transitions), they are regularly used in robotics as well [21, 22]

able to generalize over existing adaptive exploration strategies. Furthermore, we prove in the paper that ADEU can achieve optimal exploratory behaviour for a wide class of games. Finally, we put forward ADEU versions that can *safely explore* in safety-critical environments, as confirmed by preliminary results.

To test our framework, we create different instances of ADEU, each corresponding to a different uncertainty-measuring mechanism. We test our method in the most complex/difficult among the robotic tasks in [36], and in the DeepSea domain. Our experiments demonstrate that the performance of an ADEU instance depends on the ability of each uncertainty-measuring mechanism to quantify uncertainty. However, in our experiments, the instance of ADEU employing a specific uncertainty-measuring mechanism, outperforms the original exploration method that employs the same uncertainty-measuring mechanism. Moreover, there is *always* some instance of ADEU that performs better than all of its opponents. As such, ADEU can be considered as a ‘plug-and-play’ adaptive exploration framework, that can be coupled with any uncertainty-measuring mechanism; and exhibits the ability to improve the performance of DRL algorithms that utilize the corresponding uncertainty-measuring mechanism in their exploration process.

## 2 THE ADEU FRAMEWORK

*A Motivating Example:* Let us consider a variation of the classic Frozen Lake example as shown in Figure 1.<sup>2</sup> In that environment, the agent has to discover the sole trajectory that will eventually lead it to the goal, passing through frozen tiles, while if it steps on a “hole” tile the episode terminates. In our example, the agent is rewarded when it reaches the next frozen tile. This reward is inversely proportional to the distance of the goal: the reward function outputs a higher reward when the agent reaches an allowed tile that is also closer to the goal. In addition, let us assume that an agent has already been trained to reach state 61 (state at the end of the red line in Figure 1—right above state 71). In other words, its policy  $\pi(s)$  can successfully lead it to state 61, replicating the red trajectory. The agent aims to extend the red trajectory until reaching the treasure.



**Figure 1: The modified Frozen Lake environment. Exploration gets harder as dimensions increase.**

An effective way for that agent to explore this environment is by “trusting” its policy until state 61 and then switching to exploration.

<sup>2</sup>Frozen Lake’s official documentation: [https://gymnasium.farama.org/environments/toy\\_text/frozen\\_lake](https://gymnasium.farama.org/environments/toy_text/frozen_lake)

Assume an agent employing a classic exploration technique like  $\epsilon$ -greedy (or any noise-based strategy) that explores every episode with a *constant*  $\epsilon$  (or constant variance in the case of a noise-based strategy). By not modifying  $\epsilon$  during an episode, the agent explores at a constant rate. Then, the agent will probably not be able to replicate the red trajectory since it is probable (approximately  $\epsilon\%$  at each timestep) that the agent will deviate from the known policy.

In an even worse scenario, assume an exploration strategy that *forces* the agent to visit states with high uncertainty (e.g., a state that has not been visited enough times, or a state whose value the agent is unsure about). In that scenario, the agent is uncertain for all the states not belonging to the trajectory. Thus, this exploration strategy will force the agent to needlessly explore rather than trusting the trajectory, leading to the truncation of a large number of episodes before managing to actually extend the trajectory. The use of a *greedy-in-the-limit* exploration strategy would lead to similar “over-exploration” phenomena.

Now, suppose an adaptive exploration method that determines at every step whether (i) the agent should explore a state, or (ii) it should exploit/trust the known policy for that state. By appropriately alternating between these two phases, the agent can replicate the known trajectory and subsequently begin exploration. A key challenge lies in determining how the agent identifies *the appropriate moment* to transition between the phases.

Let us define  $f(s) = 1/(\beta\sqrt{n(s)})$  where  $n(s)$  is a (pseudo-)counter that counts (or approximates in larger environments) the times that a state  $s$  has already been visited, and  $\beta < 1$  a positive constant. Let us also suppose that the states of the trajectory have been visited numerous times. Then,  $f(s)$  arguably provides an effective signal on whether to switch/alternate between the two phases. This suggests an agent should explore proportionally to the value of  $f(s)$ .

### 2.1 Formal Framework Description

Against this background, we put forward ADEU, a framework allowing for adaptive exploration, based on the uncertainty associated with each state. ADEU makes use of some uncertainty-measuring mechanism  $f(s)$ , whose value dictates when an agent should start exploring and when it should trust the gained knowledge, similarly to the example. As discussed earlier, uncertainty is a subjective term; its definition is closely related to the problem being solved. The ADEU framework implicitly accepts any definition of uncertainty via the implementation of an  $f(s)$ . As such, practically all uncertainty-measuring mechanisms—e.g., those utilized by *intrinsic motivation* methods to detect novel states, or by *epistemic uncertainty*-driven exploration to quantify the convergence degree of a policy—are compatible with ADEU. We will elaborate on this after formally presenting the key ADEU principles.

The ADEU framework can be summarized concretely by Equation 1, which defines the way that an ADEU agent selects an action given a state  $s$ .

$$a \sim \mathcal{D}(\pi(s), g(f(s))) \quad (1)$$

where  $\mathcal{D}$  is the distribution of choice,  $\pi(s)$  is the current policy for a state  $s$ ,  $f(s)$  is the aforementioned uncertainty-measuring mechanism and  $g()$  is a normalizer function.

In more detail, this action will be chosen by sampling from a distribution  $D$ . The first moment of that distribution is the policy  $\pi(s)$  that the DRL agent has already learned. Hence, the *mean* of the probability distribution  $D$  (intuitively, its center of mass) is the policy  $\pi(s)$ . The *second moment (variance)* of that distribution is  $g(f(s))$ , which is proportional to the uncertainty. Hence,  $g(f(s))$  defines how “spread” the distribution is, and how close the sampled action  $a$  will be to the policy  $\pi(s)$ .

The key intuition in ADEU is that the spread of the distribution should be defined so that it corresponds to the agent’s uncertainty regarding the  $\pi(s)$  policy; while that uncertainty can be calculated in any way considered appropriate.<sup>3</sup> To this end,  $f(s)$  in  $g(f(s))$  is *some* mechanism able to compute uncertainty, like the ones used in uncertainty-based exploration methods[6, 18, 25] or mechanisms designed for intrinsic reward construction in intrinsic motivation methods[3, 5, 21].<sup>4</sup> By default,  $f(s)$  will output a positive scalar. In multidimensional distributions  $D$ , this scalar can multiply an Identity matrix. In addition, since  $f(s)$  can be *any* uncertainty-measuring mechanism, we use  $g(\cdot)$  as a normalizer to prevent  $f(s)$  from outputting enormous values. Equipped with  $f$  and  $g$ , ADEU can effectively utilize any uncertainty-measuring mechanism proposed in previous (deep exploration) works.

Moreover,  $D$  can be *any* distribution deemed appropriate. In this paper, we use the Multinomial distribution in discrete action space experiments/examples, and the Gaussian (or Normal) distribution in continuous action space ones. We selected these distributions due to their simple implementation.

Guided by Equation 1 to select actions, an ADEU agent is able to alternate between exploitation and exploration within the episode in a principled manner, and its behaviour can be summarized as follows. At every timestep, the agent (i) calculates the uncertainty  $f(s)$  regarding that state; (ii) constructs a distribution  $D$  in which the first moment is the current policy  $\pi(s)$ , and the second moment is the  $g(f(s))$ ; and (iii) samples from that distribution. In case of low uncertainty, the distribution will not be “spread” and hence the sampled action will be close to  $\pi(s)$ . Notice that this effectively causes the agent to replicate parts of a trajectory corresponding to a  $\pi$  it is confident about. In any other case, the sampled action will probably be far from  $\pi(s)$ , leading the agent to explore new actions in that state. This new experience can potentially lead to an extended trajectory.

Note that the discussion above does not imply that ADEU needs any trajectories or other forms of background knowledge. ADEU assumes a zero-length initial trajectory at the beginning of the training process, and then is progressively able to extend it via its adaptive exploration technique.<sup>5</sup> In other words, at the start of training, the agent has not yet converged at a policy, even for the initial states. As such, it lacks certainty about which actions to take

<sup>3</sup>In this paper, we use the first raw moment (mean) and the second central moment (variance), both in the context of continuous and discrete distributions. One could conceivably use any other moments of choice to enhance action selection. For instance, one could use some function of the third standardized moment in order to guide exploration towards a specific direction, e.g., a direction that is deemed *safe* [38, 39], if such a direction can be in some way associated with the kurtosis of the distribution.

<sup>4</sup>To clarify, in our method, we use the mechanism used in intrinsic motivation methods for defining  $f(s)$ , and *not* for providing an exploration bonus.

<sup>5</sup>Of course, if ADEU is provided with an accurate initial trajectory or policy, it can exploit such knowledge.

and must explore until it converges to a reliable policy, even for the initial state. Once this occurs, the agent follows the learnt action policy from the initial state until it reaches the next state where “enough” uncertainty arises. There, it resumes exploration. This exploration and policy refinement cycle continues until the agent becomes confident across the entire trajectory.

We now proceed to define a generic class of exploration-related decision problems (or exploration-related single-agent “games”) such as the “modified Frozen Lake” one, for which ADEU agents are able to achieve optimal exploratory behaviour. These are “increasing-reward” settings: intuitively, only the best action transfers the agent from state  $s$  to a “better” state  $s^*$  where  $\max_a r(s^*, a) \geq \max_a r(s, a)$ . (In our definition and discussion below,  $s$  is the current state,  $s'$  some future ones, and  $s^*$  is the best future one. We will also say that  $s$  “neighbours”  $s'$  if  $T(s, \cdot, s') > 0$ ).

**DEFINITION 2.1 (INCREASING-REWARD SINGLE-AGENT GAMES).** *In Increasing-Reward Single-Agent Games, an agent has to decide on an action  $a$  to execute at a state  $s$ , given that the following hold.*

- $\exists!(s^*, a^*) : r(s^*, a^*) \geq \max_a r(s, a)$ . *There is a unique future  $s^*, a^*$  pair, with  $s^*$  neighbouring  $s$ , in which the agent achieves a higher reward than when playing  $\arg \max_a r(s, a)$  in  $s$ .*
- $\forall s' \neq s^*, \exists a : \max_a r(s', a) < r(s, a)$ . *That is, for all future states  $s' \neq s^*$  “neighbouring”  $s$ , the agent achieves a lower maximum reward than the reward achieved in the current state  $s$  for at least one action  $a$ .*
- $\exists! b : b = \arg \max_a r(s, a), \forall s$ . *That is, at every state  $s$ , there is a unique action  $b$  that maximizes the reward function in that state  $s$ . Moreover, for that same  $b$  it holds:*
- $T(s, b, s^*) = 1$ , *where  $r(s^*, a^*) > \max_{a'} r(s', a')$  where  $s'$  neighbours  $s, s' \neq s^*$ , and  $(s^*, a^*)$  is the unique pair mentioned above—i.e., when the agent executes  $b$  in  $s$ , it will transition with probability 1 to the (unique)  $s^*$  that is (strictly) better (in terms of maximum reward) than any other neighbour state  $s'$ .*
- *Increasing-Reward Single Agent Games inherit all other characteristics from standard MDP models [30].*

**COROLLARY 2.1.** *In Increasing-Reward Single-Agent Games, the optimal behaviour for an agent is to explore a state  $s$  until action  $b : b = \arg \max_a r(s, a)$  is found; no further exploration is then required for  $s$ .*

Notice that a plethora of standard games (such as the one described in Fig. 1, and also sorting, finding the sole existing trajectory in complex maze environments, etc.) are considered as Increasing-Reward Single-Agent Games. We now prove that ADEU (equipped with an effective  $f(s)$ ) achieves efficient exploratory behaviour in such games with a large but finite action space. In more detail, assuming that after a finite amount of time, the ADEU agent will discover action  $b$  for a state  $s$ , the  $f(s)$  will decrease and stop the exploration in that specific state  $s$ . Hence, for this ADEU agent  $\pi(s) = b$  and the agent will explore the future state  $s^*$ . A suitable  $f(s)$  should measure uncertainty as, e.g., the difference between the current policy  $\pi(s)$  and  $b = \arg \max_a r(s, a)$ ; or as the  $f(s)$  used in the modified Frozen Lake

**DEFINITION 2.2.** *Following the aforementioned notation, we define as “best neighbour state at timestep  $i - 1$ ” the state  $s_i^* : T(s_{i-1}, b, s_i^*) = 1$  and  $b = \arg \max_a r(s_{i-1}, a)$*

**THEOREM 1.** *ADEU agents that have executed  $b := \arg \max_a r(s_i, a)$  for some  $s_i$  during a training episode, will reduce their exploration of  $s_i$  relative to the effectiveness of their  $f(s)$  mechanism and focus on exploring  $s_{i+1}^*$ .*

**PROOF.** The proof can be found in an extended version of this paper provided in <https://arxiv.org/abs/2509.03219>.  $\square$

Theorem 1 shows that an ADEU agent, when provided with an effective uncertainty-measuring mechanism, will solve an Increasing-Reward Single-Agent Game. The theorem can be extended to guarantee efficient (in expectation) ADEU behavior in environments with stochastic reward or transition functions also. ADEU’s efficiency arises from the fact that it does not needlessly explore, but instead replicates the parts of a “known” or “learnt” trajectory that it trusts. But this trajectory might be sub-optimal. To keep improving the learnt trajectory, ADEU also performs the so-called rollout episodes. In more detail, at the beginning of every episode, with probability  $\rho$ , ADEU performs exploration using as uncertainty mechanism the  $f_{\text{rollout}}(s) = c$ , where  $c$  is a constant number. Hence, for that *specific* “rollout” episode, ADEU takes the chance to explore so as to improve the learnt trajectory. We should note that in the case of a well-constructed  $f(s)$  that smoothly decreases,  $\rho$  should and can simply be set to 0. For simplicity, in our implementation we treat  $\rho$  as a constant that does not change throughout training. Also, we set  $f_{\text{rollout}}(s) = c$  to present a simple version of our framework, but, in more complex environments, this can change too.

Both the rollout episodes and the use of more complex definitions of uncertainty, allow the ADEU agent to effectively explore different, and more complex settings (e.g., even settings requiring deep exploration like DeepSea [26]) than the ones in Definition 2.1, as discussed immediately below and as our experiments demonstrate.

## 2.2 Generality of the ADEU framework

In this section, we elaborate on the fact that ADEU is a generic adaptive exploration framework. First, ADEU can readily use as  $f(s)$  any predefined mechanism—e.g., some epistemic uncertainty-measuring mechanism or mechanisms used in intrinsic motivation methods to approximate visitation frequency. Moreover, it is a framework that can emulate most, if not all, existing adaptive exploration techniques. This can be achieved by designing an as-complex-as-required  $f(s, \cdot)$ —e.g., an  $f(s, \cdot)$  mechanism that outputs a value considering previous states or actions, or other constraints.

*Using predefined mechanisms as  $f(s)$ :* To see this, consider first the *intrinsic motivation (IM)* family, which rewards the agent whenever a novel state is visited. This non-stationary reward, termed as  $r_{\text{intrinsic}}(s)$ , is added to the external reward provided by the environment, as described in Equation 2. Then, the agent acts greedily to maximize Equation 2.

$$r(s) = r_{\text{extrinsic}}(s) + r_{\text{intrinsic}}(s) \quad (2)$$

To calculate the  $r_{\text{intrinsic}}(s)$ , IM-oriented methods [3, 5, 21, 22] propose numerous mechanisms that can quantify the novelty of a state. For example, RND [5] proposes to calculate the visitation frequency as the approximation error between two networks named Target and Predictor. Predictor network is trained to predict the Target’s output for a given state. The difference between the output of the

two networks (Equation 3) approximates the visitation frequency for that specific state.

$$r_{\text{intrinsic}}^{\text{RND}}(s) = \|f_{\text{predictor}}(s) - f_{\text{target}}(s)\|_2^2 \quad (3)$$

ADEU can utilize this approximation error as an uncertainty-measuring mechanism. That is, an instance of ADEU that defines uncertainty as the visitation frequency of a state can utilize  $f(s) = r_{\text{intrinsic}}^{\text{RND}}(s)$ . In a similar way, any other mechanism approximating visitation frequency, such as the ones proposed in other IM methods [3, 21, 31] can also be employed.

One can also create an instance of ADEU that utilizes an uncertainty mechanism similar to the ones defined in epistemic-uncertainty driven methods [6, 18]. *Epistemic uncertainty*-driven methods employ different representations of the same  $Q$ -function to measure how certain an agent is about its policy in a specific state. For example, the UCB [6, 18] strategy greedily selects actions that maximize the upper bound of the  $Q$ -function shown in Equation 4.

$$a = \arg \max_a [Q_{\text{mean}}(s, a) + \lambda Q_{\text{std}}(s, a)] \quad (4)$$

where  $Q_{\text{mean}}(s, a)$  is the mean  $Q$ -value calculated across the different representations of the  $Q$ -function,  $Q_{\text{std}}(s, a)$  is the corresponding standard deviation, and  $\lambda$  is a positive constant. Given this, we can also create an ADEU instance that utilizes the above uncertainty-measuring mechanism by setting  $f(s) = \lambda Q_{\text{std}}(s, \pi(s))$ . Table 1 summarizes the discussion above.

**Table 1: Instances of ADEU using different uncertainty-measuring mechanisms—examples only; any IM or mechanism calculating confidence over  $\pi$  can be used.**

Definition of Uncertainty	An example of $f(s)$
Novelty of $s$	$f_{\text{RND}}(s) = \ f_{\text{predictor}}(s) - f_{\text{target}}(s)\ _2^2$
Confidence on $\pi(s)$	$f_{\text{UCB}}(s) = \lambda Q_{\text{std}}(s, \pi(s))$

*Emulating adaptive exploration techniques:* Table 2 presents certain key adaptive exploration works, and their instantiation within the ADEU framework. See also “Related Work” for more details.

Table 2 explains how ADEU incorporates previous works on adaptive exploration. For instance, ADEU can emulate Dabney et al. [8] by utilizing a Bernoulli distribution, which will “decide” over two different scenarios: (i) policy  $\pi(s)$  with probability  $1 - p$ ; and (ii) an option with probability  $p$ . The  $p$  parameter of the Bernoulli distribution will dynamically change according to  $f(s, n)$ . In more detail, the initial value for  $n$  is 0. At the first call of  $f(s, n)$ ,  $n$  will be sampled from a heavy-tail distribution and then,  $p = f(s, n)$ , where  $f(s, n) = 1$  when  $n > 0$ , and otherwise it follows the procedure described in Dabney et al. [8]. In addition, ADEU can also improve  $\epsilon z$ -greedy by incorporating more complex definitions of uncertainty, maintaining the use of the options. Hence, “extensions” of  $\epsilon z$ -greedy [4, 17], that improve Dabney et al. [8] by modifying the way that the length  $n$  of a trajectory is decided, can also be incorporated by ADEU. In addition, instead of having a Bernoulli distribution, we can utilize a Categorical one. The process of modifying  $p$  will remain the same, however, in this scenario, we can simplify the “task” of finding an option.

<sup>6</sup> $n$  is sampled in the first call to  $f(s, n)$ , then decremented in subsequent calls.

**Table 2: Previous adaptive exploration methods as special cases of ADEU.**

Description	Special Case of ADEU
Tokic [34]: $\epsilon$ -greedy where $\epsilon$ is adapted at each state based on TD error.	Utilizing a Bernoulli distribution as $D$ . Let $f(s) = \text{TD-Error}(s)$ .
Dabney et al. [8]: Executes a fixed-length option with prob. $\epsilon$ ; otherwise follows $\pi(s)$ . Length sampled from a heavy-tailed distribution.	A Bernoulli decides between $\pi(s)$ and an option. We will utilize a more complex uncertainty-measuring mechanism $f(s, n)$ , that calculates uncertainty according (also) to a parameter $n$ —this is still allowed by ADEU as previously discussed. This mechanism sets the probabilities $p$ of the Bernoulli based on $n$ , which is sampled once and decremented at each step. <sup>6</sup>
Pislar et al. [27]: Explores intra-episodically after a trigger signal, for a fixed duration.	$f(s)$ replaces the trigger. A Bernoulli distribution is utilized. $f(s)$ modifies the probability $p$ to favor exploration, then outputs a constant $p$ for a fixed number of steps.
Ecoffet et al. [10, 11], Torne Villasevil et al. [35]: The agent first returns to a state, then explores.	Use Bernoulli or Gaussian. $f(s)$ is binary and switches phases.

Given that ADEU can subsume existing adaptive exploration methods as special cases, it is reasonable to expect that it can replicate their performance in the tested domains. Thus, ADEU is well-suited for tasks with *both* dense and sparse reward structures. For instance, *ADEU solves the extremely demanding DeepSea domain [26] by emulating Dabney et al.’s [8] adaptive exploration method.*

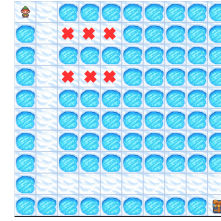
*Motivating example cont’d:* We now continue with the technical setup and the results of the motivation example. We design a large-scale grid environment of size  $1500 \times 1500$ , extending the setup shown in Figure 1. In this environment, we evaluate four different agents: (i) a standard  $Q$ -learning agent using  $\epsilon$ -greedy exploration, (ii) an intrinsic motivation (IM) agent built on top of  $Q$ -learning, (iii) a UCB-based ensemble agent composed of multiple  $Q$ -learners, and (iv) an instance of ADEU. For the IM agent, we adopt ideas from [31] without the need for approximation due to the discrete state space. The ADEU agent adjusts its action selection based on state visitation frequency, favoring exploration in less-visited states and exploiting learned values in frequently visited ones. It uses the same mechanism as the IM agent.

Table 3 shows the maximum reward and the mean reward during learning, computed as follows. We employed 5 different runs (seeds) for all agents, with each run comprising  $3 \times 10^4$  episodes. An episode terminates when: (i) the agent reaches the goal; or (ii) the agent falls in a hole; or (iii) the agent acts/survives  $t > 3 \times 1500$ . The following process was performed for each run: Once every 100 episodes, an “evaluation” episode was executed, in which the agent greedily used its policy (i.e., no actions were selected for exploration during that episode); and we stored the reward accumulated during that episode, and all such episodes in each run. We then created an “average run” by averaging over these episodes’ values across all runs; and computed the mean episodic reward and the maximum episodic reward in that average run. Note that the maximum episodic reward was achieved after convergence.

As shown in Table 3, ADEU demonstrates superior performance compared to the other methods by rapidly learning a policy that

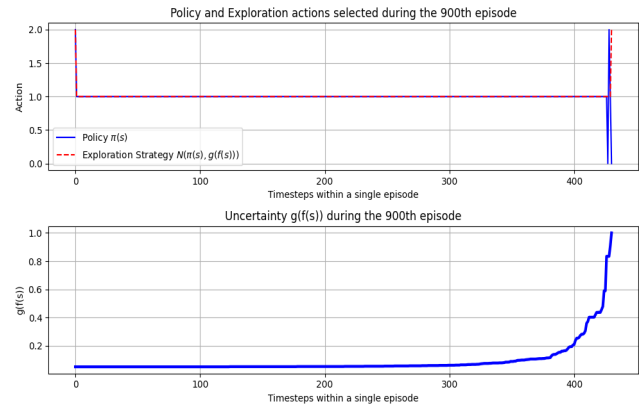
**Table 3: Mean and Max episodic rewards in the Frozen Lake domain (average over 5 runs). ADEU was built on top of a  $Q$ -learning agent.**

Exploration Strategy	Mean and Max episodic rewards
$\epsilon$ -greedy	334; 983
ADEU	88,854; 104,487
IM [31]	393; 989
UCB [6]	36,070; 104,487



**Figure 2: Environment used in testing safety. In this particular figure, the dimensions of the grid are  $10 \times 10$ . The red crosses represent the unsafe states.**

consistently guides it toward the target price. Specifically, the ADEU agent avoids excessive exploration of frequently visited states—those characterized by low uncertainty according to the employed function  $f(s)$ . Instead, it tends to replicate the learnt trajectory until completion before initiating further exploration; and manages to learn the correct trajectory. By contrast, both the  $\epsilon$ -greedy and IM algorithms fail to learn a successful trajectory to the goal, likely due to their persistent over-explorative behaviour. Finally, UCB learns the correct trajectory, but it requires nearly four times as many episodes as ADEU (16,800 vs 4,515 for ADEU). Its average reward is approximately half that of ADEU’s. Note that in our main experiments involving more complex action and state spaces, UCB always ranks behind both ADEU variants (see Table 5).



**Figure 3: (Top) Blue line is  $\pi(s)$  across one episode. Red line shows the action selected by the agent. (Bottom) Uncertainty as calculated by the agent across the states of a single episode.**

Figure 3 (top subfigure) shows both the actual policy  $\pi(s)$  and the action selected from the ADEU agent (so, using Equation 1) during a single episode in our Frozen Lake environment. In addition, the bottom subfigure shows the value of  $g(f(s))$  during the same episode. As expected, at the beginning of the episode,  $g(f(s))$  outputs low values since the agent visits states that it is certain about. Hence, the action selected by the ADEU agent coincides with that dictated by  $\pi(s)$ . As the agent visits states with high uncertainty, it does not trust its so far learned policy and starts exploring them. As of that, the action selected by ADEU and the policy  $\pi(s)$  differ at the end of the episode, since  $g(f(s))$  increases. Note also that the action suggested by  $\pi(s)$  in the later timesteps of the episode is not correct. However, the action sampled for exploration was the correct one, allowing the agent to “survive” for more timesteps and *gain useful experience*. Thus, indeed, the exploration phase should have begun at the later timesteps; had it started earlier, the agent would not have “survived” to extend the known optimal policy.<sup>7</sup>

### 2.3 ADEU for safe exploration

Now, as discussed and seen in the example, ADEU avoids exploration in states with low uncertainty. Moreover, it can be extended so that it conducts *safe exploration*—i.e., to not explore in *unsafe* states [38, 39]. By simply defining a new mechanism that also considers this extra constraint, ADEU can avoid exploring these states.

More precisely, suppose that a mentor has already informed the agent about some potential unsafe states (like in [32]) and the policy  $\pi(s)$  the agent should follow in these particular states. We would like the agent to *not* explore finding a better policy in these specific states. Hence, by designing an  $f(s)$  that outputs low uncertainty in these states, an ADEU agent can alternate between exploration and exploitation proportionally to the safety of the current state. Such an ADEU agent, will simply follow Equation 1.

Alternatively, the agent can be informed about the cost function of the environment (like in [33]). By replacing  $f(s)$  with the cost function, ADEU can combine adaptive exploration with safety. Even if the ADEU agent is *not informed* about unsafe states or the cost function, it can interact with the environment to approximate it. Along with this “approximation” process, the agent still respects the safety constraints. In that case, ADEU only requires an initial safe policy as in [2] (e.g., one identified via the approach of [40]); i.e., Eq. 5 allows for the aforementioned “safety-driven” or “safety-aware” behaviour, requiring no knowledge of the cost function.

$$a \sim \mathcal{D}(\pi(s), g(1/f_{\text{safety}}(s))) \quad (5)$$

where  $f_{\text{safety}}(s)$  is the expectation of the environment’s cost function  $\mathbb{E}_{s,a} \max_a \{g_{\text{cost}}(s, a)\}$  (where  $g_{\text{cost}}(s, a)$  follows the standard definition described in [1]), which is learned by interactions.

By sampling from Equation 5, the agent will replicate the known policy  $\pi(s)$  in a state with high expected cost; in any other state, the agent will explore randomly. In more detail, in a state with high  $f_{\text{safety}}(s)$ , the sampled action  $a$  will not deviate from  $\pi(s)$ , given the low variance in Equation 5. By contrast, in a state with low  $f_{\text{safety}}(s)$ , the sampled action will promote exploration. Finally, Equation 6 shows how we can combine Equations 1 and 5 to achieve *both* (i)

*safe exploration* when required; and (ii) *uncertainty-driven adaptive exploration* in states where safety does not matter:

$$a \sim \mathcal{D}(\pi(s), g(f_{\text{uncertainty}}(s)/\lambda f_{\text{safety}}(s))) \quad (6)$$

In Eq. 6  $f_{\text{uncertainty}}(s)$  is the aforementioned uncertainty-measuring mechanism and all other notation (except  $\lambda$ ) follows Eq. 5. Note that  $\lambda$  generally acts as a term (or mechanism) that ensures the safety term “dominates” when safety is critical. For example,  $\lambda$  can take “conditioned” values for a state  $s$  e.g.,  $\lambda = [\max_a \{Q(s, a)\}]$  if  $\mathbb{E}_{s,a} \max_a \{g_{\text{cost}}(s, a)\} > \max_a \{Q(s, a)\}$ , otherwise 1]. Therefore, the  $f_{\text{safety}}(s)$  term will be amplified when the expected cost is high.

In this case, in states with high expected cost, the agent will act according to the known policy, in other states, it will act according to uncertainty-driven adaptive exploration. Thus, Equation 6 proposes a general agent that can take the best of both worlds.

**Table 4: Performance metrics of different exploration strategies in the Safety Frozen Lake domain. Results are averages over 5 runs. ADEU was built on top of a Q-learning agent.**

Exploration Strategy	Mean & Max Reward	Mean Constr.
$\epsilon$ -greedy	210; 739	0.34
IM [31]	267; 780	0.55
UCB [6]	12,605; 65,687	0.10
ADEU (using Equation 1; $f(s)$ =Visitation freq.)	<b>59,865; 104,487</b>	0.01
ADEU (using Equation 5, i.e., safety-aware)	7;7	<b>0.00</b>
ADEU (using Equation 6; $f_{\text{uncertainty}}$ =Visitation freq.)	19,255; 85,087	<b>0.00</b>

Figure 2 shows a variant of the modified Frozen Lake that also incorporates some constraints. In more detail, the red crosses represent non-terminal states that the agent should not visit. In cases that the agent visits such states, a *safety constraint* is violated, and also the agent gets an immediate reward equal to 0. Otherwise, it is rewarded according to the previous example. In addition, we conducted experiments in a  $1500 \times 1500$  grid similarly to Section 2.2.

To test both aforementioned scenarios, we create 3 different instances of ADEU: (i) an agent that has been pretrained on a safe policy in states neighboring unsafe states, and explores using the visitation frequency definition of uncertainty (Equation 1); (ii) a safe ADEU agent that has been pretrained on the same safe policy, but explores according only to the cost function (Equation 5), which is learned simultaneously with policy  $\pi(s)$ ; and (iii) a safe ADEU agent built on top of the aforementioned one, which explores the environment according both to the learned cost function and the visitation frequency definition of uncertainty (Equation 6).

Table 4 shows the mean and maximum episodic evaluation rewards throughout the training process. In addition, it shows the mean episodic constraint violations. All ADEU instances outperform their opponents since they achieve fewer constraint violations. Since all algorithms tested in a domain requiring efficient exploration, the ADEU instance that only uses the learned cost function (Eq. 5) did not yield a high reward due to the lack of (efficient) exploration in safe states. By contrast, the ADEU instance of Eq. 1 effectively explores the environment and finds the final goal faster, but at the cost of some constraint violations—still violates *an order of magnitude fewer constraints* than its non-ADEU opponents. Finally, the ADEU instance that combines safety with uncertainty-driven adaptive exploration (Eq. 6) yields a high reward, achieving 0 constraint violations, utilizing the best of both worlds (uncertainty-driven adaptive exploration *and* safe exploration).

<sup>7</sup>We expect the same behaviour in more difficult multi-dimensional domains such as the ones described in Section 3. However, we cannot create a figure similar to Figure 3 for those, since both  $\pi(s)$ ,  $a \in \mathbb{R}^n$ , where  $n > 2$ .

**Table 5: Mean and Max episodic rewards in various domains [36]. Results are averages over 15 runs.**

Domain	TD3 + ADEU			TD3+UCB	TD3+RND	TD3+Noisy Nets
	$f(s) = c$	$f(s) = \text{RND}$	$f(s) = \lambda Q_{\text{std}}(s, \pi(s))$			
Walker2d	564; 726	<b>745; 947</b>	413; 500	355; 397	1; 23	519; 657
Hopper	376; 631	<b>733; 1167</b>	293; 514	245; 422	10; 35	429; 766
Swimmer	40; 46	<b>52; 62</b>	36; 39	32; 36	-3; -2	30; 35
Ant	345; 980	<b>1072; 1341</b>	<b>1134; 1500</b>	585; 790	-2658; 980	364; 981
Humanoid	86; 132	313; 345	<b>512; 658</b>	89; 149	85; 132	75; 88
Hum. Stand.	61511; 64490	<b>73895; 79978</b>	<b>72978; 80139</b>	54928; 56527	55211; 56057	62152; 63645

### 3 EXPERIMENTAL EVALUATION

To experimentally test our framework in earnest, we create two main instances of ADEU, using as  $f(s)$ : (i) the intrinsic motivation mechanism proposed in Burda et al. [5]; and (ii) the epistemic uncertainty measuring mechanism proposed in [6]. These are the agents described in Table 1. Both agents were built on top of a TD3 [13] DRL algorithm’s original implementation.<sup>8</sup> As such, we refer to them as: (i) TD3+ADEU with  $f(s) = \text{RND}$ ,<sup>9</sup> and (ii) TD3+ADEU with  $f(s) = \lambda Q_{\text{std}}(s, \pi(s))$ . In addition, we consider an instance of ADEU with an *uninformative*  $f(s)$ , and create a “TD3 + ADEU with  $f(s) = c$ ” agent, where  $c$  is a constant. Obviously,  $f(s) = c$  is *not* an appropriate uncertainty measuring mechanism. The agent with constant  $f(s)$  will serve as baseline, since its distribution matches the one used in TD3 [13] and hence, they share the same exploration strategy. We use  $g(f(s)) = c$  for *this* ADEU instance—while for the others, we use Eq. 7, with  $c = 0.2$  as in TD3 implementation.

$$g(f(s)) = \text{sigmoid}(f(s)) \cdot c \quad (7)$$

Our agents use TD3 as a “basis” algorithm. In principle, any DRL algorithm could have been used instead. Since ours is an exploration framework, we used a DRL algorithm not tied to any exploration strategy—and TD3 uses a ‘vanilla’ one that can be readily replaced with ADEU. By contrast, SAC [14] uses the entropy framework, and hence, it is tied to that framework’s particular exploration strategy.

We also implement three agents to act as our main competitors, to test ADEU against representatives of three standard families of exploration methods: IM exploration, epistemic uncertainty-driven exploration, and (parameter-space) noise exploration. Specifically, we test against TD3+RND, an agent that acts greedily in the environment with a reward function equal to  $r_{\text{extrinsic}}(s) + r_{\text{intrinsic}}^{\text{RND}}(s)$  (cf. Equations 2 and 3). In addition, we compare against TD3+UCB, an agent that selects actions aiming to maximize Equation 4.

Finally, for interest, we also create an agent that uses Noisy Nets [12] to perturb the actor’s network and achieve more efficient exploration. We compare against this TD3+Noisy Nets agent, since Noisy Nets is a popular choice for exploring robotic domains. Note that Noisy Nets does not utilize any definition of uncertainty—it is just an “improved” TD3 agent in terms of exploration.

We first conducted experiments in standard MuJoCo domains [36]. Specifically, we used the five domains considered the most difficult according to the official documentation.<sup>10</sup> Such domains do not satisfy Definition 2.1. However, the ADEU agents still achieve (as

later described) superior performance against their opponents. In these domains, the agent should explore the *continuous multidimensional action space* to find the best policy. A dense reward function is used to describe these robotic tasks. However, ADEU can also be used in environments with a sparse reward function, when using appropriate uncertainty-measuring mechanisms e.g., [3, 24].

Table 5 presents agents’ performance in terms of mean and max episodic accumulated reward in the aforementioned domains. These mean and maximum values were calculated over 15 runs (seeds), via the process described in the motivation example. We present both maximum and mean reward to track both the best performance of an algorithm and an indication of the time required to reach convergence. We mark in bold the best-performing algorithms in terms of mean episodic reward within each domain—an algorithm is considered to be “best-performing” if its mean episodic reward reaches at least 90% of the highest mean episodic reward accumulated.

*ADEU improves DRL performance:* As seen in Table 5, TD3+ADEU with  $f(s) = \text{RND}$  or TD3+ADEU  $f(s) = \lambda Q_{\text{std}}(s, \pi(s))$  outperform the others in *every* domain. In addition, in the three most difficult domains (i.e., Ant, Humanoid Standup, and Humanoid) *both* of those ADEU instances outperform *all* other agents, achieving the first and the second best performance. It is also noteworthy that an instance of ADEU equipped with a specific uncertainty-measuring mechanism  $f(s)$  typically outperforms the original exploration method that employs the corresponding  $f(s)$  mechanism.

In addition, our results show that ADEU with  $\lambda Q_{\text{std}}(s, \pi(s))$  outperforms most of its opponents in the three aforementioned highly challenging environments. This indicates this mechanism tracks uncertainty more efficiently in larger environments.

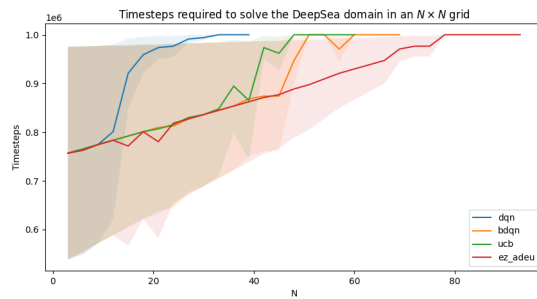
*ADEU is a ‘plug-and-play’ adaptive exploration framework:* As seen in our experiments, ADEU can incorporate any uncertainty-measuring mechanism of choice. Of course, as also discussed above, not all uncertainty-measuring mechanisms are created equal. Notice, e.g., that the *naive* instance of TD3+ADEU with  $f(s) = c$  is outperformed by at least one other ADEU instance in each domain. This indicates that ADEU efficiency depends on the efficiency of the uncertainty mechanism. Seen otherwise, the choice of an appropriate uncertainty-measuring mechanism for use in ADEU, promises the improvement of performance in the domain of interest.

*Results in the DeepSea domain:* We also employed ADEU in DeepSea, in order to put the agent’s ability for deep exploration to an arduous test. Specifically, we tested the emulation of Dabney et al. [8] from the ADEU framework, hereafter referred to as *ez-adeu*, in this demanding domain. Figure 4 shows the timesteps required for an algorithm to solve (i.e., find the tressure) a DeepSea domain

<sup>8</sup><https://github.com/sfujim/TD3>.

<sup>9</sup>More precisely but less concisely, “TD3+ADEU with  $f(s) = r_{\text{intrinsic}}^{\text{RND}}$ ”.

<sup>10</sup><https://gymnasium.farama.org/environments/mujoco/>. We did not use [9] since its domains do not have the “early termination” condition.



**Figure 4: Timesteps required for each algorithm to solve the DeepSea domain for  $N \times N$  grid. The lower the curve, the better the performance of the algorithm.**

with a  $N \times N$  grid. The  $x$ -axis of that figure shows the different values of  $N$  while the  $y$ -axis shows the aforementioned timesteps.

As seen in Figure 4,  $\epsilon z$ -adeu achieves better performance in terms of scalability in the DeepSea domain. Specifically,  $\epsilon z$ -adeu solves the DeepSea domain for  $N \approx 80$  when its best opponents struggle even for  $N > 60$ . We should also highlight that both the BDQN and UCB opponents were built on top of an ensemble of DQNs—an architecture that promotes Deep Exploration—in contrast to  $\epsilon z$ -adeu, which was built on top of a standard DQN method.

It is clear from our experiments that ADEU, paired with efficient uncertainty-measuring mechanisms, can be utilized for both demanding control tasks and domains requiring deep exploration.

## 4 RELATED WORK

We now briefly discuss related work regarding IM, epistemic uncertainty, or adaptive exploration.

*Intrinsic Motivation:* Several IM approaches have been tested in robotic domains. For instance, [21] design an efficient intrinsic motivation function that can be utilized as an uncertainty mechanism in ADEU. They test their method in a sparse reward variance of the Fetch domain; an easier domain than standard robotics. Then, [22] adds a random reward to the sparse reward function, to guide the agent to explore different parts of the environment. This can be used by ADEU in extremely sparse reward function domains.

*Epistemic Uncertainty:* As mentioned earlier, UCB [6, 18] is commonly used in continuous action-space robotic domains. We also created an instance of ADEU using this ensemble mechanism to track uncertainty (the ADEU instance with  $f(s) = \lambda Q_{\text{std}}(s, \pi(s))$  in Sec. 3) and test it against a UCB implementation built on top of a TD3 agent. Da Silva et al. [7] “asks” a mentor for guidance when the uncertainty is high; uncertainty was measured in a similar way.

*Active Exploration:* Several works [15, 16, 24, 28] have proposed methods to efficiently explore the environment to derive an  $\epsilon$ -optimal policy, while also obtaining a low sample complexity. Some methods [15, 28], where tested in tabular or “discretized” environments. Hence, these works tackle a distinct exploration problem, and there is no straightforward comparison with ADEU. Other works [16, 24] incorporate these ideas in the classic DRL testbeds either by proposing mechanisms to motivate efficient exploration [24]

or by changing the reward function that the agent aims to maximize [16]. Both mechanisms can be used in the ADEU framework.

*Adaptive exploration:* Tokic [34] uses the TD-Error to change the  $\epsilon$  of the  $\epsilon$ -greedy in each state. Hence, an agent is more likely to explore when uncertainty about the  $Q$ -function is high. Pislari et al. [27] is the first work that actively attempts to create a framework for *intra-episodic* exploration. They propose numerous external “signals” for changing modes from exploitation to exploration, either heuristic or random ones. As noted earlier, these can be used as ADEU uncertainty mechanisms. Similarly to [27],  $\epsilon z$ -greedy [8] suggests following directed behaviour with random—sampled from a heavy-tailed distribution—length. This directed behaviour guides the agent to repeat a single action numerous times. Finally, most goal-based exploration works [10, 11, 35] can be viewed as ‘adaptive exploration’ ones since they replicate the known trajectory with no exploration, and then use some exploration to extend it.

## 5 CONCLUSIONS AND FUTURE WORK

In this paper, we put forward a generic adaptive exploration framework that uses a generic uncertainty-based action selection mechanism to decide in a principled manner when to alternate between exploration and exploitation. The generality of that mechanism allows (i) our framework to effectively incorporate any uncertainty-measuring mechanism of choice; and (ii) existing adaptive exploration techniques to be viewed as special cases of our framework. In problems requiring adaptive exploration, ADEU is a promising and easily deployable solution. Additionally, it allows the user to select an existing uncertainty measurement mechanism or to define a heuristic one tailored to its problem. Our experiments verified the superiority of ADEU in difficult testbeds.

Ongoing and future work includes extending ADEU in various directions. To begin, we intend to perform tests to further verify ADEU’s ability to exploit background knowledge; and its ability to recover from being fed with incorrect background information or to detect and escape clear suboptimal trajectories by optimizing the rollout procedure. Moreover, we are actively working to extend ADEU to *guarantee* safe exploration [38, 39]. In ongoing work, we are calculating the upper bound of allowed exploration in neighboring safe states, so that ADEU can guarantee safe exploration. We also aim to extend ADEU to multiagent environments. By leveraging agents’ independent exploration, ADEU agents can adaptively explore the environment. Furthermore, by allowing a centralized agent to define  $f_i(s)$  for each agent  $i$ , we can reduce the non-stationarity issues that usually arise in multiagent RL. Finally, we intend to design mechanisms to appropriately set the third and fourth moments of  $D$ , to *guide* exploration to particular directions.

## ACKNOWLEDGMENTS

The research described in this paper was carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union - NextGenerationEU (Implementation Body: HFRI. Project name: DEEP-REBAYES. HFRI Project Number 15430).

## REFERENCES

- [1] Eitan Altman. 1999. *Constrained Markov Decision Processes: Stochastic Modeling*. Routledge.

- [2] Yarden As, Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Stelian Coros, and Andreas Krause. 2025. ActSafe: Active Exploration with Safety Constraints for Reinforcement Learning. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=aKRADWBJ11>
- [3] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskiy, Daniel Guo, Bilal Piot, Steven Kapturovski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. 2020. Never Give Up: Learning Directed Exploration Strategies. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Sye57xStvB>
- [4] André Biedenkapp, Raghu Rajan, Frank Hutter, and Marius Lindauer. 2021. Temporal: Learning when to act. In *International Conference on Machine Learning*. PMLR, 914–924.
- [5] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by random network distillation. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=H1lJnR5Ym>
- [6] Richard Y. Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. 2017. UCB Exploration via Q-Ensembles. *arXiv* (nov 2017). <https://doi.org/10.48550/arXiv.1706.01502>
- [7] F. L. Da Silva, P. Hernandez-Leal, B. Kartal, and M. E. Taylor. 2020. Uncertainty-Aware Action Advising for Deep Reinforcement Learning Agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. AAAI Press, 5792–5799. <https://doi.org/10.1609/aaai.v34i04.6036>
- [8] Will Dabney, Georg Ostrovski, and Andre Barreto. 2021. Temporally-Extended  $\epsilon$ -Greedy Exploration. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ONBPHF7zG4>
- [9] Rodrigo de Lazcano, Kallinteris Andreas, Jun Jet Tai, Seungjae Ryan Lee, and Jordan Terry. 2024. *Gymnasium Robotics*. <http://github.com/Farama-Foundation/Gymnasium-Robotics>
- [10] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. 2019. Go-Explore: a New Approach for Hard-Exploration Problems. *CoRR abs/1901.10995* (2019). [arXiv:1901.10995](https://arxiv.org/abs/1901.10995) <http://arxiv.org/abs/1901.10995>
- [11] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. 2021. First return, then explore. *Nature* 590, 7847 (feb 2021), 580–586. <https://doi.org/10.1038/s41586-020-03157-9>
- [12] Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. 2018. Noisy Networks For Exploration. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rywHCPkAW>
- [13] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*. PMLR, 1587–1596.
- [14] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *CoRR abs/1801.01290* (2018). [arXiv:1801.01290](https://arxiv.org/abs/1801.01290) <http://arxiv.org/abs/1801.01290>
- [15] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. 2019. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*. PMLR, 2681–2691.
- [16] Zhang-Wei Hong, Tsu-Jui Fu, Tzu-Yun Shann, and Chun-Yi Lee. 2020. Adversarial active exploration for inverse dynamics model learning. In *Conference on Robot Learning*. PMLR, 552–565.
- [17] Joongkyu Lee, Seung Joon Park, Yunhao Tang, and Min-hwan Oh. 2024. Learning Uncertainty-Aware Temporally-Extended Actions. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 12 (Mar. 2024), 13391–13399. <https://doi.org/10.1609/aaai.v38i12.29241>
- [18] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2021. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*. PMLR, 6131–6141.
- [19] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *ICLR, Yoshua Bengio and Yann LeCun* (Eds.). <http://dblp.uni-trier.de/db/conf/iclr/iclr2016.html#LillicrapHPHETS15>
- [20] Zhixuan Lin, Pierluca D’Oro, Evgenii Nikishin, and Aaron Courville. 2024. The Curse of Diversity in Ensemble-Based Exploration. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=M3QXCOTtk4>
- [21] Sam Lobel, Akhil Bagaria, and George Konidaris. 2023. Flipping coins to estimate pseudocounts for exploration in reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) (ICML ’23). JMLR.org, Article 939, 20 pages.
- [22] Srinath V. Mahankali, Zhang-Wei Hong, Ayush Sekhari, Alexander Rakhlin, and Pulkit Agrawal. 2024. Random Latent Exploration for Deep Reinforcement Learning. In *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=Y9qzwNIKVV>
- [23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fiedelnd, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmash Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (feb 2015), 529–533. <https://doi.org/10.1038/nature14236>
- [24] Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. 2019. Information-Directed Exploration for Deep Reinforcement Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Byx83s09Km>
- [25] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep Exploration via Bootstrapped DQN. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf)
- [26] Ian Osband, Benjamin Van Roy, Daniel J Russo, and Zheng Wen. 2019. Deep exploration via randomized value functions. *Journal of Machine Learning Research* 20, 124 (2019), 1–62.
- [27] Miruna Pslar, David Szepesvari, Georg Ostrovski, Diana L. Borsa, and Tom Schaul. 2022. When should agents explore?. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=dEwfxT14bca>
- [28] Alessio Russo and Alexandre Proutiere. 2023. Model-Free Active Exploration in Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=YEtstXJpP3>
- [29] Hassam Sheikh, Mariano Phielipp, and Ladislav Boloni. 2022. Maximizing Ensemble Diversity in Deep Reinforcement Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=hjd-kcpDp2f>
- [30] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press. <http://incompleteideas.net/book/the-book-2nd.html>
- [31] Haoran Tang, Rein Houthoofd, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. 2017. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems* 30 (2017).
- [32] Garrett Thomas, Yuping Luo, and Tengyu Ma. 2021. Safe reinforcement learning by imagining the near future. *Advances in Neural Information Processing Systems* 34 (2021), 13859–13869.
- [33] Haozhe Tian, Homayoun Hamedmoghadam, Robert Noel Shorten, and Pietro Ferraro. 2024. Reinforcement Learning with Adaptive Regularization for Safe Control of Critical Systems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=MRO2QHydPF>
- [34] Michel Tokic. 2010. Adaptive  $\epsilon$ -Greedy Exploration in Reinforcement Learning Based on Value Differences. In *KI 2010: Advances in Artificial Intelligence (Lecture Notes in Computer Science)*, Rüdiger Dillmann, Jürgen Beyerer, Uwe D. Hanebeck, and Tanja Schultz (Eds.). Springer, Berlin, Heidelberg, 203–210. [https://doi.org/10.1007/978-3-642-16111-7\\_23](https://doi.org/10.1007/978-3-642-16111-7_23)
- [35] Marcel Torme Villasevil, Max Balsells I Pamiés, Zihan Wang, Samedh Desai, Tao Chen, Pulkit Agrawal, and Abhishek Gupta. 2023. Breadcrumbs to the Goal: Goal-Conditioned Exploration from Human-in-the-Loop Feedback. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 63222–63258. [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/c7c7cf10082e454b9662a686ce6f1b6f-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/c7c7cf10082e454b9662a686ce6f1b6f-Paper-Conference.pdf)
- [36] Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo De Lazcano Perez-Vicente, J K Terry, Andrea Pierré, Sander V Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. 2025. Gymnasium: A Standard Interface for Reinforcement Learning Environments. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=qPMLvJxtPK>
- [37] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
- [38] Akifumi Wachi, Wataru Hashimoto, Xun Shen, and Kazumune Hashimoto. 2023. Safe exploration in reinforcement learning: A generalized formulation and algorithms. *Advances in Neural Information Processing Systems* 36 (2023), 29252–29272.
- [39] Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Xueqian Wang, Bo Yuan, and Dacheng Tao. 2022. Penalized Proximal Policy Optimization for Safe Reinforcement Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3744–3750. <https://doi.org/10.24963/ijcai.2022/520> Main Track.
- [40] Edoardo Zorzi, Alberto Castellini, Leonidas Bakopoulos, Georgios Chalkiadakis, and Alessandro Farinelli. 2025. Seldonian Reinforcement Learning for Ad Hoc Teamwork. *Reinforcement Learning Journal* 6 (2025), 2082–2100.