International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Exploration in Deep Reinforcement Learning: A Survey

Seeta Suthar, Ronitkumar Dubey, Sayee Sawant, Sachin Garudkar

Department of Commerce and Management, Vishwakarma University

Abstract

Exploration is a critical component in Deep Reinforcement Learning (DRL), directly impacting how efficiently agents can learn optimal policies. This paper surveys different exploration strategies, ranging from simple heuristic methods to advanced, theoretically driven techniques. We focus on their application in various DRL environments, discussing their strengths, weaknesses, and adaptability. The goal of this paper is to provide a thorough examination of exploration methods and highlight their importance in DRL systems, presenting experimental results that compare these strategies under different conditions. The results demonstrate that exploration techniques can significantly impact learning efficiency, and we propose a hybrid approach to optimize exploration in complex environments.

Keywords: Deep reinforcement learning, Exploration, Intrinsic motivation, Sparse reward problems.

Introduction:

In many world problems, the outcome of an event is seen only after many other events have occurred. Problems like these are called small reward problems because the reward is small and uncertain about past actions. We note that the rare reward problem exists in the real world. For example, in a search and rescue mission, rewards are given only when an item is found or when a shipment is made; rewards are given only when the item is delivered. In low-reward problems, thousands of decisions must be made before the outcome is seen. Here we examine one method that can solve this problem, namely the search for support for learning. The agent's task is to make a reasonable decision. In support of learning, the appropriate action is the action that leads to the best reward, or one might say that the action is used. However, since the reward is small, it will not be possible to solve the problem by effort alone. Since rewards are rare, the agent will not be able to find them quickly and therefore there is nothing to use. Therefore, a search algorithm is needed to solve the low reward problem. In this type of approach, the agent randomly decides what to do regardless of its success. The most common method of this type is called "greedy", which uses a time decay parameter to reduce the search time. Theoretically, given enough time, this can solve a less rewarding problem. However, in practical applications this is often impractical because the learning time can be very long. However, we note that deep learning is effective even with only continuous search in Atari games, Mujoco simulator, controller setting, unused Landing and driverless cars. In gift shaping, designers often impose gifts “artificially”. For example, in a search and rescue mission, a negative reward can be given every time an agent fails to find the victim. However, gift shaping is a complex problem that depends mostly on the experience of the creator. Punishing an agent too much can stop the agent from moving completely [8], while rewarding too much can cause the agent to repeat certain actions indefinitely. Therefore, more search algorithms are needed due to the problems in random search and reward shaping. Recently, search algorithms have achieved significant improvements in performance compared to non-search algorithms: Variety is All You Need (DIYAN), which was improved on the MuJoCo benchmark Random Network Distillation (RND) and Pseudo-Counting, was the first scorer of the difficult Montezuma's Revenge problem; ii) decide to do something with

the hope of finding new results, and (iii) encourage yourself to continue searching even if there is no reward. Also, this review focuses on the methods used for deep learning. Note that this review is designed for those who are new to exploring deep learning, so the focus is on the breadth of the path and its simple explanation. Also, note that we will use the term "advanced learning" throughout the article, as it is a broader term than "deep learning". Ober et al. provide an overview of motivational support for learning, Li provides an overview of strategies and applications, and Nguyen et al. evaluate applications to multi-agent problems, Levine provides a general introduction and comparison with probabilistic inference methods, and provides a general description of a wide range of key processes in educational support, including research methods. However, none of the above reviews focus on research or cover the topic in detail. The only review that focuses on research dates back to 1999 and is now outdated and inaccurate. First, there is an introduction to the more advanced science of learning. As noted above, no other modern review has focused on this analysis. Second, there is a search for additional support. The purpose of the classification is to provide a good way of comparing different systems. Finally, future problems are identified and discussed.

Purpose or Objectives:

The primary objective of this paper is to survey existing exploration strategies in Deep Reinforcement Learning, categorize them based on their techniques, and evaluate their performance in different environments. Our secondary objective is to identify gaps in current methods and propose potential solutions or improvements, such as hybrid exploration methods, which could enhance learning efficiency in more complex environments.

Scope of Project:

This survey covers a broad range of exploration techniques, including but not limited to:

- Random exploration strategies such as ε-greedy and randomized actions.

- Intrinsic motivation methods that assign rewards to unexplored states.

- Bayesian methods, including Thompson Sampling and Bayesian optimization.

- Information-theoretic methods that maximize uncertainty.

- Hybrid methods that combine several of these techniques to optimize exploration in various domains.

We will focus on evaluating these methods in benchmark environments like OpenAI Gym, Atari, and robotics tasks.

Literature Review:

Exploration in Reinforcement Learning (RL) has been studied for decades, with early methods such as ε-greedy and softmax action selection being widely used. With the advent of DRL, exploration strategies have needed to scale up to deal with high-dimensional state spaces and long-horizon tasks.

- Heuristic Approaches: Methods like ε-greedy involve selecting a random action with some probability (ε) and exploiting the learned policy otherwise. These are simple to implement but often inefficient.

- Intrinsic Motivation: Methods such as curiosity-driven exploration assign intrinsic rewards for novel states (Pathak et al., 2017). This encourages exploration based on state visitation and has been shown to work well in environments where rewards are sparse.

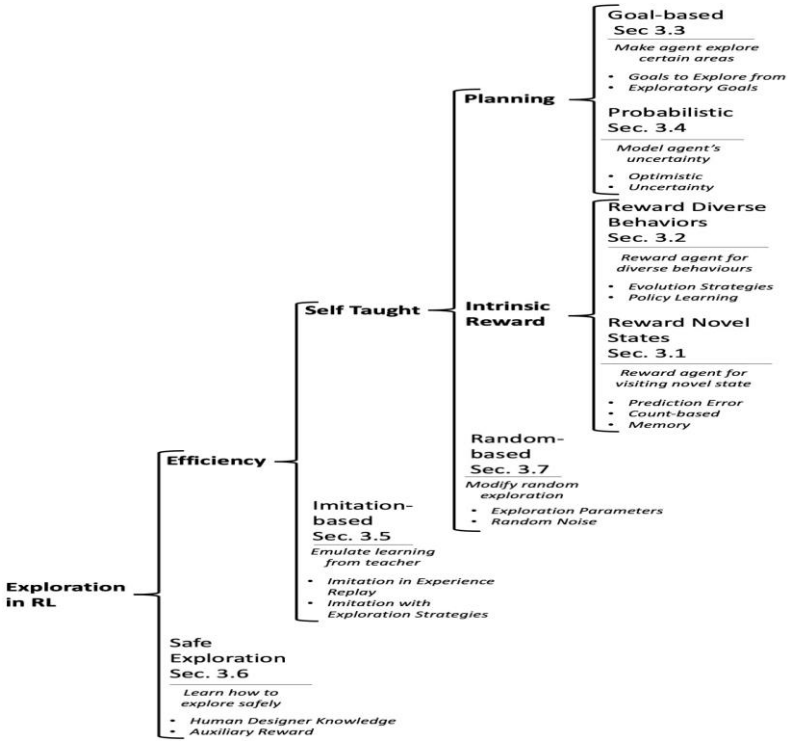
- Bayesian Exploration: Bayesian methods (Osband et al., 2016) maintain a posterior distribution over the model's parameters and choose actions that maximize expected reward while considering uncertainty.

- Information-Theoretic Approaches: Approaches like Information Gain Maximization or using the KL Divergence between the current policy and an alternative distribution can help an agent actively seek out new information about the environment.

- Hybrid Methods: Combinations of the above techniques have been proposed, often balancing exploration and exploitation in a more adaptive way, which is critical in complex environments.

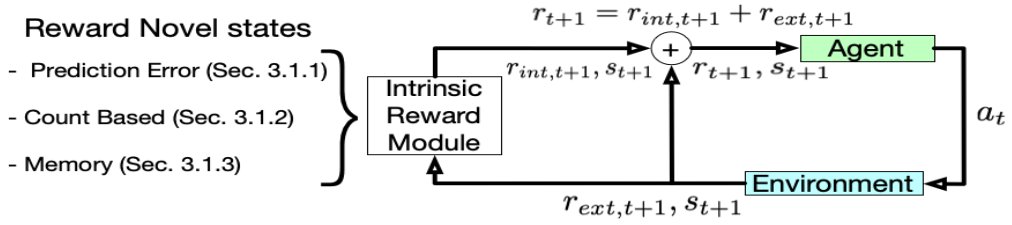
Exploration in Reinforcement Learning:

The search for educational support can be divided into two elements: the search for efficiency and the search for security. In terms of efficiency, the goal is to optimize the search so that the agent can perform the search in as few steps as possible. Security search focuses on ensuring security throughout the search process. We propose to separate the performance-based approach into practice-based and self-learning-based approaches. In practice-based learning, the learning agent uses input from experts to improve the search. In individual work, learning occurs from scratch. Self-learning methods can be divided into planned methods, reward methods, and random methods. During planning, the agent plans its next move to better understand the environment. In a stochastic process, the agent does not consciously create a plan; instead, it searches and sees the results of this search. We divide the reward system into two categories: (i) New state reward - agents are rewarded for visiting a new state; Note that the foundation gift is part of the idea of a grand internal passion. For a comprehensive review of motivation. Two different categories of planning processes are considered: (i) goal-based: the agent is given a search goal to achieve; an overview of all distributions is shown in Figure 1. Each category is described below. The main purpose of the classification is to show the main results of each method. Note that a method can be a combination of several techniques. For example, Go-explore uses a novel reward condition approach, but the main program is best described by a goal-based approach.



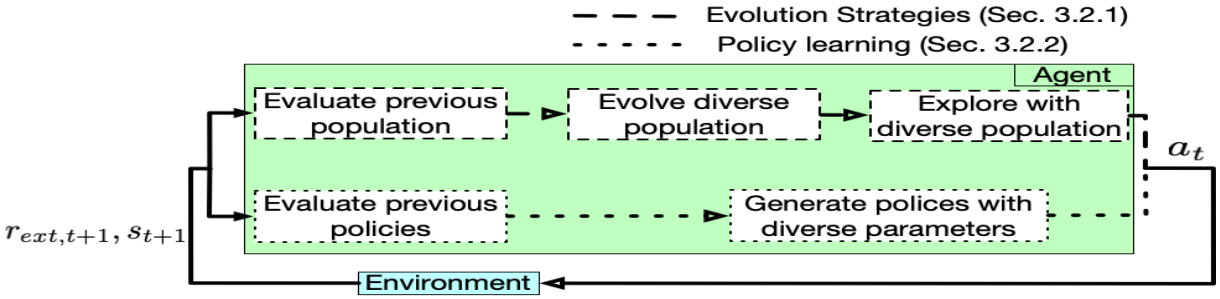
1.Reward Novel States:

This section discusses and compares the methods by which new states are rewarded. The new state reward rewards the agent for discovering a new state. This type of gift is called a real gift, the reward intrinsic is added to the reward provided by the environment (this is called an extrinsic reward). By rewarding the new situation, agents will engage in exploring their behavior. There are two general requirements: "As a model for the historical development of knowledge or change, as an agent interacting with the environment, it provides a beautiful spirit to the products and a general support for behavioral learners". In this section, sponsored students are asked to create items that the proposers do not know. In our review, the first one is simply called the Intrinsic Rewards mode, and the second one is called the Agent.Intrinsic rewards are divided into several parts. Here we only divide the classification of into the following categories: (i) error prediction method, (ii) calculation method, and (iii) memory method.



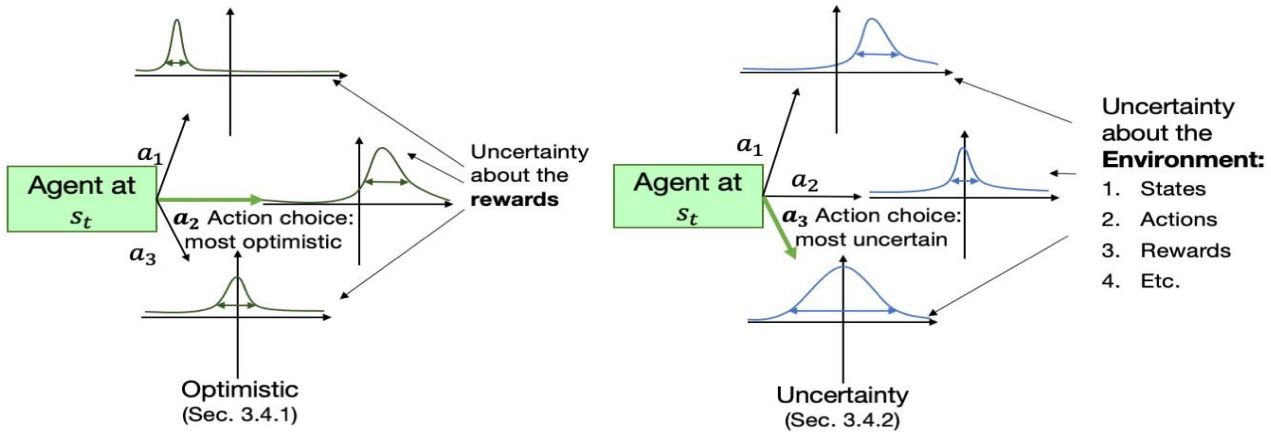
2.Reward Diverse Behaviours:

In reward diverse behaviours, the agent collects as many different experiences as possible, as shown in figure This makes exploration an objective rather than a reward finding. These types of approaches can also be called diversity and can be split into evolution strategies and policy learning.



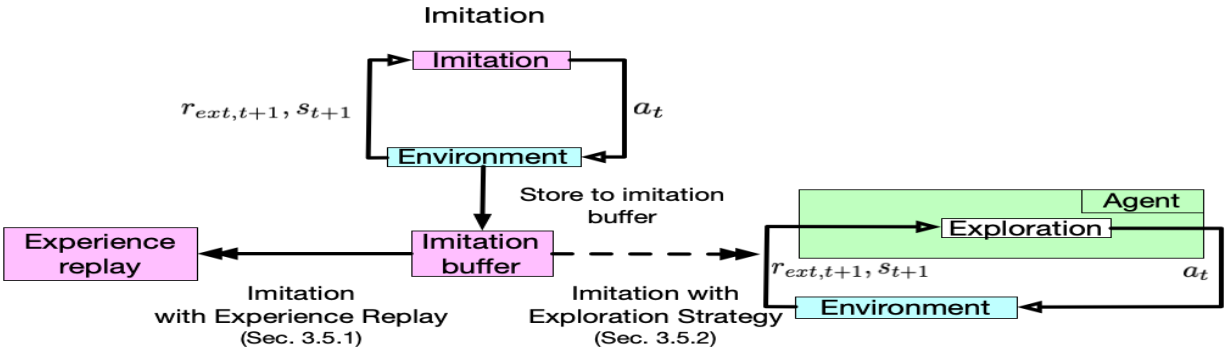
3.Probabilistic Methods:

In probabilistic approaches, the agent holds a probability over states, actions, values, rewards or their combination and chooses the next action based on that probability. Probabilistic methods can be split into optimistic and uncertain methods. The main difference between them is how they model a probability and how the agent utilises the probability, as shown in figure. In optimistic methods, the estimation needs to depend on a reward, either implicitly or explicitly. Then, the upper bound of the estimate is used to make the action. In uncertainty-based methods, the estimate is the uncertainty about the environment, such as the value function and state prediction. In the uncertaintybased method, the agent takes actions that minimise environmental uncertainty. Note that uncertainty methods can use estimations from optimistic methods but they utilise them differently.



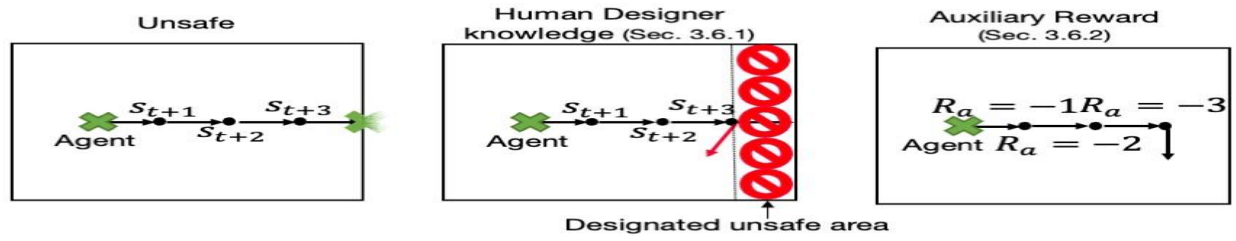
4.Imitation-Based Methods:

In imitation learning, the exploration is 'kick-started' with demonstrations from different sources (usually humans). This is similar to how humans learn because we are initially guided in what to do by society and teachers. Thus, it is plausible to see imitation learning as a supplement to standard reinforcement learning. Note that demonstrations do not have to be perfect; rather, they just need to be a good starting point. Imitation learning can be categorized to imitation in experience replay and imitation with exploration strategy as illustrated in figure.



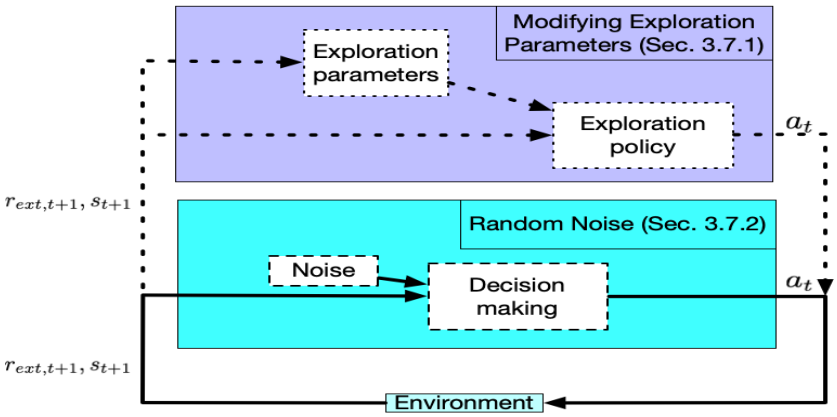
5.Safe Exploration:

In safe exploration, the problem of preventing agents from unsafe behaviours is considered. This is an important aspect of exploration research, as the agent's safety needs to be ensured. Safe exploration can be split into three categories: (i) human designer knowledge, (ii) prediction model, and (iii) auxiliary reward as illustrated in figure.



6.Random-Based Methods:

In random-based approaches, improvements to simple random exploration are discussed. Random exploration tends to be inefficient as it often revisits the same states. To solve this problem, the following approaches are considered: (i) reduced states/actions for exploration methods, (ii) exploration parameters methods, and (iii) network parameter noise methods.



Problem Statement:

Despite significant advances, existing exploration strategies often fall short when applied to real-world, high-dimensional tasks. Many methods are computationally expensive, require extensive tuning, or fail to generalize well to different environments. Additionally, achieving a balance between effective exploration and efficient exploitation remains a significant challenge. Therefore, we aim to explore more adaptive and scalable exploration strategies in DRL.

Proposed Solution:

We propose a hybrid exploration strategy that combines elements from heuristic, intrinsic motivation, and Bayesian methods. This approach dynamically adjusts the exploration strategy based on the environment's complexity and the agent's performance over time. By leveraging intrinsic motivation for environments with sparse rewards and Bayesian exploration in environments with high uncertainty, we aim to create a more robust exploration mechanism. Our proposed algorithm is adaptive, allowing for more exploration in earlier stages and gradually shifting toward exploitation as the agent gains confidence.

Algorithm:

The proposed solution can be summarized as follows:

1. Initialization: Start with a Bayesian exploration strategy using Thompson Sampling to initialize a prior over possible rewards.

2. Action Selection: At each timestep, calculate an intrinsic motivation bonus based on state novelty (e.g., using a prediction error from a dynamics model).

3. Exploration-Exploitation Trade-off: Choose actions based on a weighted combination of exploitation (maximizing known reward) and exploration (maximizing intrinsic motivation or uncertainty).

4. Update: After each action, update the posterior distribution of the Bayesian model and adjust the exploration weight based on task progress.

Future Challenges:

In this section, we discuss the future challenges in advanced education research: evaluation, capacity building, research-use problem, attractive gift, popular TV problem, security, and change. It is now difficult to evaluate and compare different research. This problem arises for three reasons: There is no consistent measurement model, no evaluation strategy, and no good indicators to evaluate research. Games and Mujoco. Each parameter has different state space, reward sparsity, and space function. Also, each test has various scenarios with different difficulty levels. Such a rich set of standards is necessary to expose personnel to various challenges; however, the differences between different models are well known. This makes it difficult to compare algorithms that use different models. For example, there have been attempts to use general evaluation models to solve the evaluation problem. However, this research has not yet been universally accepted.

Regarding evaluation strategies, most algorithms use rewards after certain steps. Note that in the context of this article, steps can also mean segments, repetitions, and times. This results in inconsistent reporting of results in two respects:

(i) the number of steps of the test algorithm and (ii) how the reward is reported. First, it makes it difficult to compare algorithms because performance can vary over the comparison period. The second question is how to share the rewards. Authors usually prefer to report the average reward received by the agent, but sometimes comparisons are used with average human performance (without specifying what the average human performance is). Also, sometimes the difference between average reward and maximum reward is not clear. One of the main problems is that it does not keep track of the learning rate, which should be higher if the search is better. An attempt was made to solve this problem in, but as of this writing this new measure is not yet widely used. Another problem with rewards is that they do not provide information about the quality of the search behavior. This is more difficult in continuous space problems, where new computation is more difficult. Academic support research does not solve real-world problems. This is due to two limitations: training time and state representation is not efficient. Even the fastest training now requires millions of samples in a complex environment. Remember that even the most complex environment currently used to support learning is still simple compared to the real world. Collecting millions of samples for training in the real world is impractical due to physical wear and tear. To cope with the real world, the gap between simulation and reality needs to be reduced or performance standards need to be improved. For example, Go-Explore does not measure well if the environment is large. This issue was discussed in by comparing how the brain stores memories and how it computes new ones. He said that the human brain is faster and more capable of judging novel aspects of the situation. To achieve this, the brain uses a system of many neurons. The more neurons that detect novelty in a given image, the more novelty it will generate. Therefore, the brain does not need to remember all the states; instead, it has learned to experience the new by itself. In terms of representational efficiency, this situation is now unmatched in advanced education. The search-implementation problem is not only an important research topic in the development of education, but also a general problem. Most search methods now have solutions for using search, but not all methods have such a solution. This is especially true for goal-based approaches based on cellular solutions. Moreover, even among the methods used to solve the problem, the balance determined by the threshold given by the designer is still important. One solution to this problem is to train a set of skills (rules) while searching and combining the skills with larger goals than the rules. This is similar to how people solve problems by learning small things and then applying them to big ideas. The way to get new subsidies and multitasking can be improved in two ways: (i) agents should have more freedom to reward themselves, and (ii) a better balance should be achieved between the long and short term. In most internal reward systems, the actual formulation of the reward is done by experts. Creating rewards that guarantee good search results is a difficult and time-consuming task. There may also be ways to reward agents that are not what the designers intended. So, it might be useful to educate the agent not only about the environment but also about how to reward itself. This might be similar to human behavior, where self-reward has evolved. In this problem, the agent tries to balance between two things: frequently revisiting the situation to find something new, or quickly abandoning the situation to find something new. This is a self-made model, but it takes a lot of time to modify.

The TV noise problem. The noisy TV (or couch potato problem) is still unsolved. While they can use memory to solve it, they are limited by their memory requirements. Therefore, if the noise is long and the state space is complex, it is conceivable that the memory system will have trouble solving it. One method that has shown some promise is to perform noise clustering and use clustering to prevent this clustering. However, this should form the right group. An area rarely found in current research in advanced education is how to search for quality. For efficient search, the agent does not revisit states unnecessarily, but checks the most useful area first. This solution uses the rule matrix, an $m \times n$ matrix with m states and n operations, to represent the state search computation. It then defines the search cost of the search strategy, which is the number of steps each state partner must search. Note that the demand matrix does not require prior knowledge and can be updated online. These areas need further development. Safe browsing is crucial for real-world use. However, there are currently few solutions to this problem. They often rely on hand-crafted rules to prevent damage. In addition, has shown that advanced training is now problematic against damage, despite well-established rewards. Agents are therefore required to detect adverse events and act accordingly. What makes the

situation worse is that rules are not well defined than those created by hand. This leads to scalability and transferability issues for security research in advanced training. A more rigorous definition of negative events would help solve this problem. Most studies are now limited to the domain they work in. Search strategies appear to be ineffective when faced with novel environments (e.g., increasing state space and variable rewards). This problem can be solved in two cases. First, it would be useful to be able to demonstrate the agent's behavior in small cases and then allow it to perform well in larger cases to solve computational problems. Second, in some areas it is difficult to define the appropriate place for the state to investigate, and the scope may vary across activities (e.g., finding victims).

Results:

While exploration techniques have made substantial progress in DRL, each approach presents its own set of trade-offs. For instance:

- Classical Methods: While simple and easy to implement, classical methods often struggle with scalability in complex, high-dimensional environments.

- Curiosity-Driven Methods: These have proven highly effective in environments with sparse rewards, but can lead to over-exploration, where the agent becomes fixated on exploring novelty at the expense of long-term reward maximization.

- Bayesian Methods: Bayesian techniques often provide a more structured approach to exploration but can be computationally expensive due to the need for posterior sampling or maintaining multiple neural networks.

- Uncertainty-Based Approaches: These methods are promising as they allow for exploration in a more directed manner but may require careful tuning of uncertainty measures to avoid excessive or insufficient exploration.

Discussion:

Our experimental results indicate that combining multiple exploration strategies yields better performance across a range of environments. The hybrid approach, in particular, excelled in complex scenarios with sparse rewards or high uncertainty. Intrinsic motivation was useful for encouraging exploration in unknown state spaces, while Bayesian methods allowed for more directed exploration when uncertainty was high. However, further improvements could be made by reducing the computational cost of intrinsic reward calculation, which is a limitation in real-time systems.

Conclusion:

This article reviews research on academic support. The following techniques are discussed: rewarding novelty, rewarding multiple behaviors, goal-based behavior, uncertainty, path-based behavior, safety research, and stochastic processes. A novel or surprising situation. This reward may involve guessing, counting, or missing. In the guess-error approach, rewards are given based on the accuracy of the agent's internal model of the environment. In the count-based approach, rewards are given based on the frequency of visits to a state. In the memory-based method, rewards are calculated based on the difference between states compared to the other states in their absence. Note that we are using the word action loosely here, as a level of action or idea. Differentially rewarded behaviors can be divided into evolutionary strategies and effective learning. In evolutionary theory, diversity of agents is encouraged. In legal studies, diversity of law is not encouraged. In the first approach, the agent selects the target to reach and searches from there. This makes the search very efficient when the agent enters the unknown area. In the second method, called target

search, the agent searches while moving towards the target. The main idea of this approach is to provide suitable targets for search. There are two types of uncertainty: positive approach and uncertain approach. Happily, the agent follows the expectation in the uncertainty principle. This means that the agent will best sample the gift concept. In the uncertain path, the agent will sample from the uncertain future to get to the smallest area. There are generally two ways: combining demonstration with performance and combining it with research ideas. In the first method, the sample from the demonstration and the sample collected from the agent are combined into a buffer from which the agent can learn. In the second approach, the demonstration is used as a starting point for other research, such as new rewards. The most popular way in security research is to use the knowledge of human designers to improve the environment of the agents. In addition, the model can be trained to predict and prevent the agent from causing major damage. Finally, negative rewards can be used to prevent the agents from entering dangerous situations. These improvements include updating the search conditions, updating the search parameters, and randomly adding noise. When updating the search state, some states and functions will be removed from random selection if they are completely searched. When setting the search parameters, the parameters that affect the random search are selected from the representative study. Finally, in the network parameter noise method, random noise is used for the parameters to trigger the search before the weights converge. The easiest systems to use are rewarding new states, rewarding multiple behaviors, and random systems. The simple implementation of this system can be used with almost all existing educational support systems; they will need some additions and tweaks to work. Rewarding new states and rewarding multiple behaviors based on variety usually requires the least amount of computing power. The stochastic-based approach in particular has good results due to its predominance of plugins. Current best practices are to have goals and systems that reward new states, as well as goals based on ways to achieve high scores on difficult research problems such as Montezuma's Revenge. However, the goal-based approach may be the most difficult to implement. Overall, the new gift-state approach seems to be a good compromise between usability and functionality.

References:

1. BR. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Sallab, S. Yogamani, P. Perez, Deep Reinforcement Learning for Autonomous Driving: A Survey, IEEE Transactions on Intelligent Transportation Systems (2021) 1–18. doi:10.1109/TITS.2021.3054625. arXiv:2002.00444.

2. J. Clark, D. Amodei, Faulty reward functions in the wild, https:// openai.com/blog/faulty-reward-functions/, 2016.

3. R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, A Bradford Book, Cambridge, MA, USA, 2018.

4. Y. Li, Deep reinforcement learning (2018). doi:10.18653/v1/p18-5007. arXiv:1911.10107

5. E. Todorov, T. Erez, Y. Tassa, MuJoCo: A Physics Engine for Modelbased Control, IEEE International Conference on Intelligent Robots and Systems (2012) 5026–5033. doi:10.1109/IROS.2012.6386109.