_2_historical_context__and__evolution_of_exploration_research _3_motivation_for_effective_exploration _2_model-based_planning__and__experience_replay _3_early_explicit_exploration_bonuses _2_bayesian_approaches_to_exploration _3_information-theoretic_exploration _2_count-based__and__density-based_novelty _3_prediction_error__and__self-supervised_curiosity _4_robust_intrinsic_rewards:_addressing_the_noisy_tv_problem _2_learning_exploration_policies_(meta-exploration) _3_integrated__and__adaptive_explora _4_population-based__and__evolutionary_exploration _2_expert-guided__and__demonstration-based_exploration _3_exploration_in_dynamic__and__ex _4_safety-aware_exploration _2_open_challenges__and__theoretical_gaps _3_emerging_trends__and__ethical_considerations

A Comprehensive Literature Review with Self-Reflection

Literature Review

October 8, 2025

Abstract

This literature review provides a comprehensive analysis of recent research in the field. The review synthesizes findings from 240 research papers, identifying key themes, methodological approaches, and future research directions.

Contents

1 Introduction to Exploration in Reinforcement Learning

1.1 The Exploration-Exploitation Dilemma

The exploration-exploitation dilemma represents a foundational and ubiquitous challenge in Reinforcement Learning (RL), fundamentally shaping how an autonomous agent acquires knowledge and optimizes its behavior within an uncertain environment (**?**). At its core, this dilemma encapsulates the inherent tension between two conflicting objectives: an agent must judiciously decide whether to *exploit* its current understanding to select actions that are known to yield high immediate rewards, or to *explore* unknown actions and states, which, despite immediate uncertainty, may lead to the discovery of significantly greater long-term rewards (**? ?**). This delicate balance is critical for the design of effective RL agents, as an imbalanced trade-off can profoundly impact learning efficiency and the ultimate optimality of the learned policy.

To formally illustrate this core dilemma, consider the classic multi-armed bandit (MAB) problem, a simplified yet powerful model for sequential decision-making under uncertainty (**?**). In a MAB setting, an agent faces K distinct "arms," each associated with an unknown probability distribution over rewards. At each time step t, the agent selects an arm $a\_t \in \{1, \ldots, K\}$ and observes a reward $r\_t \sim \mathcal{D}\_a\_t$. The objective is to maximize the cumulative reward over a sequence of T pulls, or equivalently, to minimize "regret." Regret, formally defined as the difference between the expected cumulative reward of an optimal policy (always pulling the best arm) and the agent's actual cumulative reward, is given by $R\_T = \sum \_t = 1^T(\mu$

The fundamental challenge of balancing exploration and exploitation, where an agent must gather sufficient information about its environment to make optimal decisions while simultaneously leveraging its current knowledge, has been a cornerstone of Reinforcement Learning (RL) since its inception (**?**). The historical trajectory of exploration research reflects a continuous effort to overcome the inherent complexities of unknown en-

vironments, evolving from rudimentary heuristics in simplified settings to sophisticated, scalable strategies for complex, high-dimensional domains. This evolution has been driven by both conceptual shifts and technological advancements, particularly the rise of deep learning.

The intellectual origins of principled exploration can be traced to the multi-armed bandit (MAB) problem, the simplest setting where the exploration-exploitation dilemma is starkly presented. In MABs, an agent chooses from a set of actions (arms) with unknown reward distributions, aiming to maximize cumulative reward. Foundational algorithms like Upper Confidence Bound (UCB) (**?**) emerged from the principle of "optimism in the face of uncertainty" (OFU), which encourages agents to explore actions whose true values are uncertain by optimistically assuming they might yield high rewards. Concurrently, Bayesian approaches, notably Thompson Sampling (**?**), provided a probabilistic framework for exploration by sampling from a posterior distribution over action values, effectively balancing exploration and exploitation by favoring actions with high potential given current uncertainty. These early MAB solutions laid the theoretical bedrock for later exploration strategies in full Markov Decision Processes (MDPs) (227).

Transitioning to full MDPs, early RL exploration strategies were often heuristic. The ϵ-greedy policy, a direct extension of MAB ideas, randomly selects actions with a small probability (ϵ) to discover new state-action values, while otherwise exploiting current knowledge (**?**). While simple, its undirected nature proved inefficient in larger state spaces. To address this, early research in tabular settings introduced explicit exploration bonuses, often count-based, which incentivized agents to visit less-frequented states or take less-tried actions by augmenting the reward signal. These methods aimed for broader state space coverage, but their direct reliance on explicit state-action enumeration rendered them impractical for environments with continuous or very large discrete state spaces, foreshadowing the pervasive "curse of dimensionality." The role of dynamic programming in this era was primarily to compute optimal policies *given* a known model, highlighting the critical need for effective exploration to *learn* such models in unknown environments.

A significant conceptual shift emerged with the development of theoretically grounded

exploration methods for finite MDPs, aiming to provide provable guarantees on learning efficiency. The OFU principle, originating from MABs, became a cornerstone, leading to algorithms like UCRL2 (**?**). Within the PAC-MDP (Probably Approximately Correct-MDP) framework, these methods provided provable bounds on the number of samples required to learn a near-optimal policy. While offering robust theoretical guarantees on sample complexity, these approaches were computationally demanding and inherently limited by their reliance on explicit state-action enumeration, making them largely inapplicable to the high-dimensional problems prevalent in modern RL. Efforts continued to refine these theoretical methods, with works like UCRL3 (73) introducing tighter concentration inequalities and adaptive computation of transition supports to improve practical efficiency within the theoretical paradigm. However, the fundamental trade-off persisted: rigorous theoretical guarantees often came at the cost of scalability, necessitating a paradigm shift for complex, real-world domains.

The advent of deep learning provided a new impetus for exploration research, shifting the focus towards scalable solutions for complex, high-dimensional observation spaces where traditional counting or explicit model-learning became intractable. This era saw the emergence of intrinsic motivation, a paradigm where agents generate internal reward signals for novel or surprising experiences, independent of external task rewards. The challenge was to generalize the notion of "visitation count" or "novelty" to continuous, high-dimensional state spaces. Breakthroughs included the concept of pseudo-counts (**?**), derived from density models, which allowed agents to quantify novelty in high-dimensional state spaces. Complementing this, (27) introduced ϕ-pseudocounts, generalizing state visit-counts by exploiting the same feature representation used for value function approximation, thereby rewarding exploration in feature space rather than the untransformed state space. Similarly, (**?**) demonstrated that even a simple generalization of classic count-based methods, using hash codes to count state occurrences, could achieve competitive performance in deep RL benchmarks, underscoring the power of novelty-seeking in complex environments.

Despite their success, early intrinsic motivation methods faced challenges, such as

the "noisy TV problem," where agents might be perpetually distracted by uncontrollable stochastic elements that offer no meaningful learning. To address this, (**?**) introduced the Intrinsic Curiosity Module (ICM), which generates intrinsic rewards based on the agent's prediction error of its own actions' consequences in a learned feature space, thereby focusing exploration on controllable and learnable aspects of the environment. Further refining this, (**?**) proposed Random Network Distillation (RND), a simpler and more robust intrinsic reward mechanism that measures prediction error between a policy network and a fixed, randomly initialized target network. RND proved less susceptible to environmental stochasticity, providing a more reliable signal for true novelty and significantly improving exploration stability. Simultaneously, efforts were made to bridge the gap between theoretical rigor and deep RL scalability. For instance, (43) introduced Information-Directed Sampling (IDS) for deep Q-learning, providing a tractable approximation that explicitly accounts for both parametric uncertainty and heteroscedastic observation noise, further enhancing the theoretical grounding of exploration in deep RL.

In conclusion, the evolution of exploration research in RL reflects a continuous effort to overcome the inherent challenges of unknown environments. This journey moved from foundational theoretical guarantees in simplified settings like MABs, through heuristic and theoretically-grounded methods for tabular MDPs, and ultimately to practical, scalable, and increasingly robust solutions for complex, high-dimensional domains enabled by deep learning. This historical trajectory, marked by shifts from heuristic to theoretically grounded, and then to intrinsically motivated and uncertainty-aware deep learning approaches, sets the stage for the detailed methodological discussions of advanced and adaptive strategies that follow.

1.2 Motivation for Effective Exploration

Effective exploration stands as a cornerstone for the successful application of Reinforcement Learning (RL) agents, particularly in the intricate and often unforgiving landscape of real-world environments. Its crucial role stems from the inherent challenges that frequently impede learning: the scarcity of informative reward signals, the vastness of high-

dimensional state and action spaces, the prevalence of deceptive local optima that can trap agents in suboptimal behaviors, and the critical need for policies that generalize beyond training data while remaining sample-efficient. Without well-designed exploration strategies, RL agents risk converging to inferior policies, failing to discover optimal solutions, or even remaining inert in complex tasks, thereby underscoring the continuous drive for innovation in this research domain.

One of the primary motivations for robust exploration arises from the pervasive issue of **sparse reward signals** and the **curse of dimensionality**. In many practical scenarios, agents receive meaningful feedback only after achieving specific, often distant, goals. This sparsity makes naive trial-and-error exploration highly inefficient or even impossible. Early attempts to address this, such as model-based planning with Dyna-Q (**?**) and subsequent works (**?** **?**), aimed to improve sample efficiency by leveraging learned environmental models to generate synthetic experiences. Similarly, count-based methods (**?**) offered explicit incentives for visiting less-known states. However, these foundational approaches often struggled to scale to high-dimensional or continuous state spaces, where explicit state enumeration or precise model learning becomes intractable. This limitation fundamentally motivated the development of **intrinsic motivation** techniques, which empower agents to generate their own internal reward signals, independent of external task rewards. Pioneering ideas of "curiosity" based on prediction error (**?** **?**) and learning progress (**?**) provided conceptual breakthroughs. These concepts were subsequently scaled to deep RL through methods like pseudo-counts for high-dimensional spaces (**?**), exploration bonuses derived from deep predictive models (5), and hash-based count methods (2). Such advancements have proven vital for tasks requiring extensive discovery in visually rich or complex environments, such as mapless navigation for mobile robots (29).

Beyond simply finding rewards, effective exploration is essential to overcome the **peril of deceptive local optima**. Many environments present reward landscapes with numerous suboptimal peaks, where a greedy agent might become trapped, never discovering the globally optimal policy. This challenge necessitates exploration strategies that

actively encourage agents to venture beyond seemingly good but ultimately inferior solutions. Information-theoretic approaches, such as Variational Information Maximizing Exploration (VIME) (**?**), address this by guiding agents to states that maximize information gain about the environment's dynamics, thereby reducing uncertainty and facilitating escape from local traps. More recent intrinsic motivation methods, like the Intrinsic Curiosity Module (ICM) (**?**) and Random Network Distillation (RND) (**?**), provide robust novelty signals by rewarding prediction errors in learned feature spaces or against random targets. These methods are crucial for preventing agents from being perpetually attracted to uninformative stochastic elements (the "noisy TV" problem) that could otherwise lead to spurious curiosity and suboptimal convergence. Furthermore, approaches like diversity-driven exploration (20) and novelty-seeking in evolutionary strategies (9) explicitly aim to prevent policies from being trapped in local optima by encouraging a wide range of behaviors and exploring diverse solution spaces. Robust policy optimization techniques, such as Robust Policy Optimization (RPO) (93), also contribute by maintaining sufficient policy entropy, ensuring continuous and broad exploration to avoid premature convergence.

The imperative for **sample efficiency** and **generalization** further underscores the critical need for sophisticated exploration. In real-world applications, data collection can be costly, time-consuming, or even risky, making inefficient exploration a significant bottleneck. Moreover, agents must often perform reliably in environments that differ subtly or significantly from their training conditions. This motivates exploration strategies that not only discover optimal policies quickly but also acquire knowledge transferable to unseen scenarios. For instance, (98) highlights that simple policy entropy maximization is often insufficient for sample-efficient continuous control, advocating for decoupled exploration and exploitation policies. Leveraging existing data, such as expert demonstrations (1) or large volumes of offline trajectories (125), can dramatically accelerate learning by guiding exploration towards promising regions of the state-action space, thus improving sample efficiency. The importance of exploration for *generalization* itself is a key motivation for methods like EDE (Exploration via Distributional Ensemble) (139), which encourages exploration of states with high epistemic uncertainty to acquire knowledge

that aids decision-making in novel environments. Meta-learning exploration strategies (6) enable agents to learn *how* to explore effectively across a distribution of tasks, fostering rapid adaptation and generalization. Crucially, in safety-critical domains, exploration must be conducted within predefined safe boundaries or with learned recovery mechanisms, as explored by Recovery RL (7), ensuring that the pursuit of optimal behavior does not lead to catastrophic outcomes.

In conclusion, the motivation for effective exploration in Reinforcement Learning is deeply rooted in the fundamental challenges of the field. It is indispensable for navigating sparse reward landscapes, conquering high-dimensional complexities, escaping deceptive local optima, and achieving both sample efficiency and robust generalization in dynamic, real-world settings. The continuous evolution of exploration strategies, from basic heuristics to advanced intrinsic motivation, diversity-driven methods, and meta-learning approaches, reflects its non-negotiable status as a core component for unlocking the full potential of intelligent agents. Addressing these challenges drives ongoing research to develop more robust, theoretically grounded, and computationally efficient exploration methods that can seamlessly integrate with the demands of practical applications.

2 Foundational Concepts and Early Approaches to Exploration

2.1 Basic Exploration Heuristics

The fundamental challenge of exploration in reinforcement learning (RL) lies in efficiently discovering optimal policies within an environment while simultaneously exploiting currently known good actions. The earliest and most straightforward attempts to address this exploration-exploitation dilemma centered around simple, yet widely adopted, heuristics, primarily the ϵ-greedy policy. This approach serves as a crucial baseline from which more sophisticated and targeted exploration strategies have evolved, highlighting the initial attempts to balance this fundamental trade-off.

The ϵ-greedy policy operates on a simple principle: with a small probability ϵ (epsilon), the agent selects an action uniformly at random, thereby exploring the environment. With a higher probability of $1 - \epsilon$, the agent chooses the action that maximizes its current estimated value (the greedy action), thus exploiting its learned knowledge. This method's appeal lies in its conceptual simplicity and ease of implementation, making it a foundational component in many early RL algorithms (13; 34). It ensures that every action has a non-zero probability of being selected, preventing the agent from getting permanently stuck in suboptimal policies.

Despite its widespread use, basic ϵ-greedy exploration suffers from several inherent limitations that have motivated the development of more advanced techniques. A primary drawback is its undirected nature (153). The random actions taken during exploration are not guided by any sense of novelty, uncertainty, or potential for high reward. This leads to inefficient exploration, particularly in environments with large state-action spaces, sparse rewards, or deceptive local optima (5; 51). For instance, in complex domains requiring processing raw pixel inputs, simple ϵ-greedy methods are often impractical due to their reliance on enumerating or uniformly sampling a vast, high-dimensional state-action space (5). The lack of direction means the agent might spend considerable time revisiting well-understood states or exploring unpromising regions, rather than focusing on truly unknown or potentially rewarding areas (76).

Furthermore, ϵ-greedy policies struggle to distinguish between actions that are truly unknown and those that are known to be suboptimal (6). Every action, regardless of how much is known about its outcomes, receives the same random exploration probability. This uniform randomness can be particularly problematic in real-world settings, where "random exploration, nevertheless, can result in disastrous outcomes and surprising performance" (156). The method fails to leverage the agent's uncertainty about its value estimates, a critical piece of information for efficient exploration. This limitation highlighted the need for strategies that could generalize uncertainty and direct exploration towards states or actions with high epistemic uncertainty, as explored by methods like count-based exploration in feature space (27; 2). These later approaches aimed to provide

exploration bonuses based on how frequently states or features were visited, offering a more nuanced way to encourage novelty than simple uniform random action selection.

The inefficiency of ϵ-greedy becomes even more pronounced in large or continuous state spaces, where the probability of revisiting any specific state becomes infinitesimally small, rendering simple visit counts ineffective (2). This "curse of dimensionality" necessitated methods that could generalize exploration across similar states or learn representations of novelty, moving beyond the direct, uninformative randomness of ϵ-greedy.

However, the concept of ϵ-greedy has not been entirely abandoned. Its simplicity has made it a foundational element that has been significantly refined and adapted. For example, in the context of Incremental Reinforcement Learning, where state and action spaces continually expand, a Dual-Adaptive ϵ-greedy Exploration (DAE) method has been proposed (192). This advanced variant dynamically adjusts the exploration probability ϵ based on the convergence of value estimates for specific states (Meta Policy) and guides the agent to prioritize "least-tried" actions (Explorer). This evolution demonstrates how the core idea of balancing exploration and exploitation, first introduced by basic ϵ-greedy, can be made significantly more sophisticated and targeted to address the challenges of dynamic and expanding environments, moving beyond its initial undirected and inefficient form.

In conclusion, while basic exploration heuristics like ϵ-greedy provided a crucial initial framework for addressing the exploration-exploitation trade-off, their inherent limitations—undirected exploration, inefficiency in large state spaces, and inability to distinguish between truly unknown and well-understood but suboptimal actions—underscored the necessity for more sophisticated and targeted exploration strategies. These early methods laid the groundwork, serving as a fundamental baseline from which the rich and diverse landscape of modern exploration techniques has emerged.

2.2 Model-Based Planning and Experience Replay

Efficiently navigating and learning within complex environments is a fundamental challenge in reinforcement learning (RL), often exacerbated by the high cost of real-world

interactions. Model-based planning and experience replay address this by making more efficient use of collected experience, implicitly aiding exploration by accelerating learning and propagating information more widely across the state space.

A foundational approach in this domain is the Dyna architecture, introduced by (?). Dyna-Q integrates direct reinforcement learning with planning by concurrently learning an environmental model (transitions and rewards) from real experiences. This learned model is then used to generate simulated experiences, allowing the agent to perform "mental rehearsals" and update its value function from both real and imagined interactions. This process significantly accelerates value function updates and propagates information more widely, making each real interaction more valuable and implicitly encouraging exploration by quickly refining the agent's understanding of the environment. Complementing this, (?) highlighted the importance of experience replay, a mechanism where past experiences are stored and re-used for learning. By replaying previously collected data, agents can learn more efficiently from a fixed set of interactions, reducing the need for extensive new exploration and improving sample efficiency, particularly in off-policy learning settings.

To further enhance the efficiency of model-based planning, (?) and (?) introduced prioritized sweeping. This method refines the planning process by prioritizing updates to state-action pairs whose values are likely to change significantly, or which have a large impact on other states. By focusing computational resources on the most informative simulated experiences, prioritized sweeping dramatically accelerates learning and value propagation compared to uniform sweeping, ensuring that the agent's understanding of the environment is refined more quickly and effectively. Addressing the challenge of scaling model-based methods to larger state spaces, (?) proposed reinforcement learning with a hierarchy of abstract models. This approach leverages structural decomposition to manage complexity, allowing exploration and planning to occur at different levels of temporal abstraction, which can make the learning problem more tractable. Similarly, (?) explored methods for blending planning and direct reinforcement learning, demonstrating how a learned model can be actively used to guide exploration by evaluating hypothetical scenarios and informing action selection, thereby making exploration more directed and

less random.

Beyond direct model learning for planning, advancements in state representation also contribute to the efficiency of model-based approaches. (**?**) introduced the successor representation, which models the expected future state occupancies rather than immediate transitions. While not a direct planning mechanism in the Dyna sense, this representation provides a more generalized understanding of state relationships, improving generalization for temporal difference learning and implicitly aiding exploration by making value estimates more robust and transferable across similar states.

In more recent deep reinforcement learning contexts, the principles of model-based planning continue to evolve. (5) demonstrated how deep predictive models can be used to incentivize exploration by assigning exploration bonuses based on the uncertainty or novelty derived from the learned dynamics. This approach leverages the representational power of neural networks to build scalable models in high-dimensional domains, guiding exploration towards areas where the model's predictions are less confident. Extending this, (90) proposed a model-based lifelong reinforcement learning approach that estimates a hierarchical Bayesian posterior to distill common structures across tasks. By combining this learned posterior with Bayesian exploration, their method significantly increases the sample efficiency of learning across related tasks, showcasing how sophisticated model learning can facilitate principled exploration and transfer. Furthermore, (129) integrated neural network models into an actor-critic architecture (ModelPPO) for AUV path-following control. Their neural network model learns the state transition function, allowing the agent to explore spatio-temporal patterns and achieve superior performance compared to traditional model predictive control and other RL methods, underscoring the enduring utility of learned models in complex control tasks.

Despite their significant advantages in sample efficiency and information propagation, model-based planning and experience replay methods face critical limitations. Their performance heavily relies on the accuracy and learnability of the environmental model. In complex, high-dimensional, or non-stationary domains, learning an accurate and robust model can be exceedingly challenging, and errors in the model can compound, leading to

"model bias" and potentially suboptimal policies or misleading exploration. Nevertheless, these foundational and evolving model-based approaches remain crucial for accelerating learning and making efficient use of collected experience, thereby implicitly guiding agents towards more effective exploration strategies.

2.3 Early Explicit Exploration Bonuses

The fundamental challenge of exploration in reinforcement learning (RL) necessitates strategies that transcend purely random actions to efficiently discover optimal policies, particularly in environments characterized by sparse or delayed rewards. This subsection focuses on the pioneering methods that introduced explicit incentives for agents to explore novel or less-visited states, thereby laying the groundwork for more sophisticated intrinsic motivation techniques. These early approaches were crucial in demonstrating the power of directed exploration beyond mere stochasticity.

A seminal contribution to explicit exploration bonuses came from (**?**), who introduced count-based exploration. In this paradigm, agents receive an additional, intrinsic reward for visiting states or taking actions less frequently encountered. The core idea is straightforward: by incentivizing novelty based on visitation frequency, the agent is directly encouraged to explore uncharted regions of the state space. This approach effectively transforms the problem of undirected search into a directed quest for new experiences, ensuring broader state space coverage in tabular settings. This principle aligns with the broader concept of "optimism in the face of uncertainty," where less-known options are optimistically valued higher to encourage their selection (**?**). Such count-based mechanisms share conceptual roots with strategies employed in the multi-armed bandit problem, where algorithms like Upper Confidence Bound (UCB) leverage visitation counts (or estimates of uncertainty) to balance exploitation of known good options with exploration of less-tried ones, thereby providing a theoretical basis for directed exploration in simpler settings.

Complementing frequency-based methods, the concept of *recency-based* exploration also emerged as a valuable heuristic in early RL. While less formally enshrined in a

single seminal work compared to count-based methods, the underlying idea was to grant exploration bonuses based on the time elapsed since a state was last visited, or to prioritize states that were recently discovered but not yet thoroughly explored. These approaches aimed to prevent agents from getting stuck in local optima by encouraging them to refresh their knowledge about "stale" or neglected parts of the environment. For instance, an agent might receive a bonus inversely proportional to the number of timesteps since its last visit to a particular state, ensuring that even frequently visited states are eventually re-explored if they haven't been seen for a while. Both count-based and recency-based methods, while distinct in their temporal focus, shared the common goal of directing exploration by explicitly rewarding the agent for interacting with less familiar parts of the environment, moving beyond the undirected nature of ϵ-greedy exploration.

Despite their groundbreaking nature and effectiveness in controlled, tabular, or low-dimensional environments, these early explicit exploration bonuses faced significant limitations, primarily due to the curse of dimensionality. Count-based methods, by their very definition, require maintaining an accurate count for each unique state-action pair. In environments with large, continuous, or high-dimensional state spaces (e.g., visual observations from images), enumerating and tracking every distinct state becomes computationally infeasible and memory-prohibitive. The notion of a "unique state" itself becomes ill-defined in continuous spaces, making direct counting impossible. Similarly, recency-based methods also struggle in such complex settings, as tracking the last visit time for an astronomically large or continuous state space is equally impractical. The inability of these foundational explicit bonuses to scale effectively to real-world complexity underscored the need for more generalized and robust intrinsic motivation techniques that could approximate novelty in high-dimensional settings without explicit state enumeration.

In conclusion, early explicit exploration bonuses, encompassing both count-based (frequency) and recency-based heuristics, provided a critical foundation for directed exploration in reinforcement learning. They successfully demonstrated the power of incentivizing novelty to overcome the limitations of purely random search in environments where states could be distinctly enumerated. However, their inherent reliance on explicit state

representations severely limited their applicability to tabular or low-dimensional environments. These fundamental scalability challenges, driven by the curse of dimensionality, highlighted the need for two distinct paths forward: firstly, the development of theoretically grounded approaches that could offer provable guarantees on learning efficiency in tractable domains (as discussed in Section 3); and secondly, the creation of more advanced intrinsic motivation techniques that could generalize the notion of a "count" or "novelty" to complex, high-dimensional domains without explicit state enumeration (as will be explored in Section 4.2).

3 Theoretically Grounded Exploration Strategies

3.1 Optimism in the Face of Uncertainty (OFU) and PAC-MDP

The principle of "optimism in the face of uncertainty" (OFU) stands as a foundational pillar for theoretically grounded exploration in reinforcement learning. This paradigm dictates that when an agent faces uncertainty about the true value of a state-action pair, it should optimistically assume the highest possible reward, thereby actively incentivizing exploration of unknown or poorly understood regions of the environment. This inherent bias towards unexplored options ensures that the agent gathers sufficient data to accurately estimate values, ultimately facilitating convergence to an optimal policy. OFU is intrinsically linked to the concept of Probably Approximately Correct (PAC-MDP) guarantees, which provide strong theoretical assurances that an agent can learn a near-optimal policy with high probability within a polynomial number of interactions (**?**).

The historical development of OFU principles can be traced from the simpler multi-armed bandit (MAB) setting to full Markov Decision Processes (MDPs). In MABs, Upper Confidence Bound (UCB) algorithms, such as UCB1 (**?**), exemplify OFU by selecting actions that maximize an upper confidence bound on their estimated value. This bound is typically a sum of the empirical mean reward and a bonus term that scales with the uncertainty (e.g., inversely proportional to the square root of the number of times the arm has been pulled). This strategy ensures that arms with potentially high, but uncertain,

returns are sufficiently explored.

Extending this principle to the more complex MDP setting, algorithms like R-Max (**?**) and UCRL (Upper Confidence Reinforcement Learning) (**?**) operationalize OFU to provide PAC-MDP guarantees. R-Max constructs an explicit model of the MDP and, for any state-action pair that has not been sampled a sufficient number of times, it optimistically assigns a maximal reward (R_max) and models a self-loop transition. This design effectively "forces" the planning algorithm to prioritize exploration of these unknown regions, as they appear maximally rewarding. Once a state-action pair has been visited enough times, its estimated reward and transition dynamics are considered reliable, and the optimism is removed. Similarly, UCRL algorithms maintain confidence intervals over the estimated transition probabilities and reward functions of the MDP. At each step, UCRL computes an "optimistic" MDP whose parameters lie within these confidence intervals and whose optimal policy yields the highest possible value. The agent then acts optimally with respect to this optimistic model, ensuring that actions leading to uncertain but potentially high-reward outcomes are chosen. Another notable algorithm, UCB-Value Iteration (UCB-VI), also leverages confidence bounds on value functions to guide optimistic exploration (**?**).

These OFU-based algorithms are celebrated for their robust theoretical bounds on sample complexity, guaranteeing that an agent will find an ϵ-optimal policy (a policy whose value is within ϵ of the optimal value) within a number of interactions that scales polynomially with the size of the state space, action space, and the desired accuracy. This makes them a strong foundation for efficient learning in environments where such guarantees are paramount. However, a critical limitation arises from their reliance on explicit state enumeration and accurate model estimation, which becomes intractable in high-dimensional or continuous state spaces. As highlighted by (5), while "Bayesian and PAC-MDP approaches to the exploration problem offer strong formal guarantees," they often become "impractical in higher dimensions due to their reliance on enumerating the state-action space." This "curse of dimensionality" severely restricts their direct applicability to complex, real-world environments, a challenge reinforced by comprehensive

surveys on deep reinforcement learning exploration (17).

Despite these scalability challenges, the core tenets of OFU continue to inform contemporary research. Modern model-based RL algorithms still strive for similar guarantees, even if they employ approximations to handle larger state spaces. For instance, (81) introduces PC-MLP, a model-based RL algorithm that aims for polynomial sample complexity in both Kernelized Nonlinear Regulators and linear MDPs, demonstrating that the pursuit of theoretically efficient exploration remains active. This work, like its predecessors, relies on a planning oracle, a common assumption in algorithms with strong theoretical bounds. Furthermore, recent work by (198) explores optimistic exploration using symbolic model estimates, showcasing how OFU principles can be adapted to structured environments where symbolic representations can mitigate some of the dimensionality issues, thereby making optimistic planning more tractable.

In conclusion, the principle of optimism in the face of uncertainty, coupled with PAC-MDP guarantees, provides a robust theoretical framework for efficient exploration in reinforcement learning. These methods offer strong bounds on sample complexity and ensure convergence to near-optimal policies by systematically exploring uncertain but potentially rewarding avenues. However, their inherent reliance on explicit model construction and finite state-action spaces limits their direct applicability to the vast, high-dimensional environments common in modern deep RL. This fundamental trade-off between theoretical rigor and practical scalability has motivated the development of alternative exploration strategies, such as intrinsic motivation and approximate methods, which often sacrifice explicit PAC-MDP assurances for greater applicability in complex domains.

3.2 Bayesian Approaches to Exploration

Bayesian approaches offer a principled and theoretically grounded framework for tackling the exploration-exploitation dilemma in reinforcement learning by explicitly quantifying and managing uncertainty. These methods maintain a posterior distribution over possible models of the environment, value functions, or policies, leveraging this uncertainty to guide decision-making. The fundamental premise is that actions are not merely chosen

based on their immediate expected reward, but also for their potential to reduce epistemic uncertainty, thereby leading to more informed and efficient learning over the long term. This subsection explores key techniques, from foundational posterior sampling methods like Thompson Sampling to scalable approximations using deep ensembles and Monte Carlo dropout, highlighting their mechanisms for balancing exploration and exploitation.

Historically, the concept of Bayesian reinforcement learning dates back to early theoretical works, where agents would explicitly maintain a posterior over the entire Markov Decision Process (MDP) parameters (**?**). While providing strong theoretical guarantees, the computational intractability of maintaining and updating exact posterior distributions, especially in high-dimensional state and action spaces or with complex, non-linear dynamics models common in deep reinforcement learning (DRL), severely limited their practical application. Exact Bayesian inference often requires complex computations over continuous or high-dimensional parameter spaces, making it prohibitive for real-world scenarios.

A cornerstone technique that exemplifies the Bayesian principle is Thompson Sampling. It operates by sampling a model (or a Q-function, or a policy) from the current posterior distribution and then acting optimally with respect to that sampled entity for a period. This mechanism inherently balances exploration and exploitation: models that are highly uncertain or have not been sufficiently explored are more likely to be sampled, leading to exploration, while well-understood models guide exploitation. The elegance of Thompson Sampling lies in its ability to implicitly direct exploration towards promising yet uncertain areas. Recent advancements have focused on making Thompson Sampling more scalable and provably efficient for DRL. For instance, ishfaq20235fo present a scalable Thompson Sampling strategy for RL that directly samples the Q-function from its posterior distribution using Langevin Monte Carlo, an efficient Markov Chain Monte Carlo (MCMC) method. This approach bypasses the need for restrictive Gaussian approximations, offering a more accurate representation of the posterior and demonstrating a regret bound of $\tilde{O}(d^{3/2}H^{3/2}\sqrt{T})$ in linear Markov Decision Processes (MDPs), making it deployable in deep RL with standard optimizers. Building on this, ishfaq20245to further enhance

randomized exploration for RL by proposing an algorithmic framework that incorporates various approximate sampling methods with the computationally challenging Feel-Good Thompson Sampling (FGTS) approach. Their work yields improved regret bounds for linear MDPs and shows significant empirical gains in challenging deep exploration tasks within the Atari 57 suite, underscoring the potential of efficient approximate sampling to unlock the power of Thompson Sampling in complex environments.

Given the challenges of exact Bayesian inference, much research in DRL has focused on practical approximations for estimating epistemic uncertainty, which is crucial for effective Bayesian exploration. Deep ensembles have emerged as a prominent and effective method. By training multiple neural networks with different random initializations or data subsets, the disagreement among their predictions can serve as a proxy for epistemic uncertainty. This disagreement can then be used to generate intrinsic rewards, encouraging the agent to explore states where the ensemble's predictions diverge significantly. jiang2023qmw propose Exploration via Distributional Ensemble (EDE), a method that encourages exploration of states with high epistemic uncertainty through an ensemble of Q-value distributions. EDE demonstrates state-of-the-art performance on benchmarks like Procgen and Crafter, highlighting the importance of exploration for generalization and the efficacy of ensemble-based uncertainty. Similarly, yang2022mx5 introduce Ensemble Proximal Policy Optimization (EPPO), which learns ensemble policies and incorporates a diversity enhancement regularization over the policy space. This regularization helps to generalize to unseen states and promotes exploration by encouraging the ensemble members to maintain diverse behaviors, thereby covering a broader range of the state-action space. In safety-critical applications, zhang2024ppn leverage deep ensembles to estimate epistemic uncertainty within a safe reinforcement learning framework. Their Uncertainty-augmented Lagrangian (Lag-U) algorithm uses this uncertainty to encourage exploration and adaptively modify safety constraints, enabling a better trade-off between efficiency and risk avoidance in autonomous driving.

Another practical method for approximating Bayesian uncertainty in deep neural networks is Monte Carlo dropout. By applying dropout during inference, multiple forward

passes can be performed to obtain a distribution of predictions, from which uncertainty (e.g., variance) can be estimated. This technique provides a computationally efficient way to quantify epistemic uncertainty without training multiple separate models. wu2021r67 utilize a practical and effective dropout-based uncertainty estimation method in their Uncertainty Weighted Actor-Critic (UWAC) algorithm. While primarily applied to offline reinforcement learning to detect and down-weight out-of-distribution state-action pairs, the underlying principle of using dropout to estimate uncertainty is directly applicable to guiding exploration in online settings by incentivizing visits to states where uncertainty is high.

Despite their theoretical elegance and principled approach, a common limitation of explicit Bayesian methods remains the computational complexity associated with maintaining and updating posterior distributions. While modern approximations like MCMC, deep ensembles, and Monte Carlo dropout significantly improve scalability, they introduce their own trade-offs. Deep ensembles require training and maintaining multiple neural networks, which can be computationally expensive and memory-intensive. Monte Carlo dropout, while efficient, relies on specific assumptions about the network architecture and may not always accurately capture the true posterior uncertainty. The accuracy of these approximations directly impacts the effectiveness of the exploration strategy and the theoretical guarantees. Future research continues to focus on developing more scalable, computationally efficient, and theoretically robust Bayesian approximations that can harness the full potential of uncertainty-driven exploration in complex, high-dimensional, and real-world reinforcement learning scenarios, moving beyond heuristic exploration towards more informed and adaptive learning.

3.3 Information-Theoretic Exploration

Information-theoretic exploration strategies offer a principled framework for addressing the exploration-exploitation dilemma in Reinforcement Learning (RL) by explicitly quantifying and maximizing the expected information gain. Unlike heuristic or purely novelty-seeking approaches, these methods guide agents towards experiences that are most likely

to reduce uncertainty about the environment's dynamics, the optimal policy, or the value function. This section delineates various facets of information-theoretic exploration, emphasizing how they provide a sophisticated understanding of learning progress and model improvement.

A fundamental concept in this domain is the maximization of mutual information. A seminal work, Variational Information Maximizing Exploration (VIME) (**?**), exemplifies this by proposing an intrinsic reward signal derived from the mutual information between the agent's actions and the learned parameters of its environment dynamics model. VIME leverages variational inference to estimate this information gain, thereby incentivizing the agent to take actions that maximally reduce its uncertainty about how the environment functions. This approach moves beyond simple state visitation counts, actively driving the agent to improve its internal model of the world by seeking out states and actions that are most informative for model learning. The strength of VIME lies in its explicit link between exploration and model improvement, but its computational complexity, particularly in estimating mutual information and maintaining accurate posterior distributions over model parameters in high-dimensional settings, can be a significant challenge.

Beyond uncertainty about the environment model, information-theoretic approaches also focus on reducing uncertainty about the optimal policy or value function. Information-Directed Sampling (IDS) is a prominent example, which explicitly quantifies the value of information by maximizing the "information ratio" (**?**). This ratio balances the expected reduction in regret (or increase in reward) from gaining information against the cost of exploration. Unlike Thompson Sampling (which samples a policy from a posterior and acts greedily), IDS directly optimizes for the value of information, making it a more explicit information-theoretic strategy. While initially developed for bandit problems, IDS has been extended to Deep RL, as demonstrated by (43). This work proposes a tractable approximation of IDS for deep Q-learning, which explicitly accounts for both parametric uncertainty and heteroscedastic observation noise. By leveraging distributional reinforcement learning, this approach provides a robust exploration strategy that is particularly effective in environments with varying levels of uncertainty, outperforming

traditional methods that struggle with non-uniform return variability. The application of IDS also extends beyond traditional RL control tasks, as seen in (49), where it was used in density-based structural topology optimization to efficiently direct the search towards optimal designs by maximizing the expected value of information in generative design problems.

Another significant information-theoretic concept is empowerment, which defines an intrinsic reward as the channel capacity between an agent's actions and its future states (**? ?**). Maximizing empowerment encourages an agent to explore states where it has greater control or influence over its future, effectively driving it towards regions of the state space that offer more diverse and controllable outcomes. This perspective aligns with the broader idea of intrinsic motivation, where agents are driven by an innate desire to understand and control their environment. As highlighted by (34), information theory provides a rich taxonomy for intrinsic motivation, encompassing concepts like surprise (reduction in predictive uncertainty), novelty (information gain about unfamiliar states), and skill-learning (maximizing control over future states, i.e., empowerment). These different facets underscore how information-theoretic principles can be applied to various aspects of learning and exploration.

While Bayesian methods, discussed in Section 3.2, inherently align with information-theoretic principles by maintaining and reducing uncertainty, information-theoretic exploration distinguishes itself by explicitly formulating exploration as an optimization problem over information gain. For instance, Thompson Sampling implicitly reduces uncertainty by sampling from a posterior, but IDS or VIME directly compute or approximate the value of information. The computational overhead of precisely quantifying mutual information or channel capacity remains a primary challenge for information-theoretic methods, especially in complex, high-dimensional, and non-stationary environments. Approximations, such as those used in VIME or the Deep RL extension of IDS (43), are crucial for scalability.

In conclusion, information-theoretic exploration offers a powerful and principled lens through which to design effective exploration strategies. By explicitly valuing information

gain—whether about the environment's dynamics (VIME), the optimal policy (IDS), or the agent's control over its future (empowerment)—these methods move beyond simple heuristics to foster truly intelligent and directed discovery. Despite challenges related to computational tractability and the accurate estimation of information-theoretic quantities in complex settings, advancements in approximation techniques continue to enhance their practical applicability. Future research will likely focus on developing more efficient and robust approximations for information gain, potentially integrating with meta-learning to adaptively select optimal information-seeking strategies, and further exploring their utility in multi-agent and open-ended learning scenarios.

4 Intrinsic Motivation: Novelty, Curiosity, and Prediction Error

4.1 Early Concepts of Intrinsic Curiosity

The challenge of exploration in reinforcement learning, particularly in environments characterized by sparse or delayed extrinsic rewards, led to the development of intrinsic motivation. This paradigm shift moved beyond solely relying on external reward signals, proposing that agents could be driven by an internal 'curiosity' or 'novelty' derived from their own learning progress or model improvement. These foundational concepts laid the theoretical and conceptual groundwork for later, more sophisticated curiosity-driven and novelty-seeking exploration methods.

One of the earliest proponents of intrinsic curiosity was Schmidhuber1997, who introduced the idea of rewarding an agent for improving its world model's predictive accuracy. The agent is intrinsically motivated to explore states where its current model makes inaccurate predictions, thus seeking out "surprising" observations to reduce its uncertainty and improve its understanding of the environment. Building on this, Singh2004 further formalized the notion of intrinsic motivation, comparing and contrasting different intrinsic signals such as novelty (unfamiliarity) and surprise (prediction error), providing a more

theoretical framework for these internal drives.

As reinforcement learning moved towards more complex, high-dimensional domains, the challenge became scaling these intrinsic curiosity concepts. Stadie2015 addressed this by proposing an exploration method that assigned bonuses from a concurrently learned deep predictive model of the system dynamics. This work demonstrated how the early ideas of prediction-error-based curiosity could be extended to tasks requiring raw pixel inputs, like Atari games, by leveraging deep neural networks to parameterize the world model. Further refining the theoretical underpinnings, Houthooft2016 introduced Variational Information Maximizing Exploration (VIME), a principled, Bayesian approach that encourages agents to explore by maximizing the information gain about the environment's dynamics model. This method provides a more formal way to quantify and reduce epistemic uncertainty, guiding exploration towards states that are most informative for improving the agent's internal model.

Despite these advancements, prediction-error-based curiosity methods faced a challenge known as the "noisy TV problem," where agents could be perpetually distracted by unlearnable stochastic elements in the environment that constantly generated high prediction errors. To address this, Pathak2017 proposed the Intrinsic Curiosity Module (ICM), which computes intrinsic rewards based on the prediction error of future states in a *learned feature space* rather than raw pixel space. By learning a feature representation that is invariant to factors beyond the agent's control, ICM effectively filters out unlearnable stochasticity, allowing curiosity to focus on aspects of the environment that the agent can influence. Burda2018 further simplified and improved the robustness of curiosity-driven exploration with Random Network Distillation (RND). RND measures novelty as the prediction error of a fixed, randomly initialized target network's output by a trained prediction network, providing a highly effective intrinsic reward signal that is largely immune to the noisy TV problem because the prediction target is independent of the environment's true dynamics.

The practical utility of these curiosity-driven approaches has been demonstrated across various applications. For instance, Li2019tj1 showed how a simplified Intrinsic Curios-

ity Module (S-ICM) could be effectively integrated with off-policy reinforcement learning methods, significantly improving exploration efficiency and learning performance in robotic manipulation tasks with sparse rewards. Similarly, Zhelo2018wi8 applied curiosity-driven exploration to mapless navigation for mobile robots, validating its crucial role in improving deep reinforcement learning performance in tasks with challenging exploration requirements and enhancing generalization capabilities in unseen environments. More recently, Sun2022ul9 utilized a similarity-based curiosity module to enable aggressive quadrotor flights, demonstrating how intrinsic motivation can accelerate training and improve the robustness of policies in complex control tasks.

In conclusion, the early concepts of intrinsic curiosity marked a fundamental shift in reinforcement learning, moving from external reward dependence to internal drives based on predictability, surprise, and learning progress. These pioneering ideas, from Schmidhuber1997's initial formulation of prediction error as a motivator to the more robust and scalable deep learning-driven methods like ICM Pathak2017 and RND Burda2018, have provided effective solutions for exploration in sparse-reward environments. While significant progress has been made in making these methods robust to irrelevant stochasticity, ongoing research continues to explore how to design intrinsic reward functions that consistently align with efficient and meaningful exploration across diverse, open-ended domains, and how to balance these internal drives with external task objectives.

4.2 Count-Based and Density-Based Novelty

Effective exploration is paramount in reinforcement learning, particularly when agents operate in environments characterized by sparse extrinsic rewards or vast, high-dimensional state spaces. A prominent class of intrinsic motivation methods addresses this by quantifying the 'novelty' or 'unvisitedness' of states, generating internal reward signals that encourage agents to venture into less-frequented regions. This approach aims to foster broad state space coverage, which is often crucial for discovering optimal policies.

The foundational concept of count-based exploration, as pioneered by (**?**), involves assigning an intrinsic bonus to states inversely proportional to their visitation frequency.

In tabular or low-dimensional discrete environments, these methods provide a theoretically sound mechanism for directed exploration, ensuring that agents sufficiently explore all reachable states. However, as discussed in Section 2.3, traditional count-based approaches face a critical limitation: the curse of dimensionality. In high-dimensional or continuous state spaces, the probability of revisiting any exact state becomes infinitesimally small. This renders direct state counting impractical, as most states are encountered only once, leading to uniformly high novelty bonuses that fail to guide exploration effectively.

To bridge this gap and enable count-based exploration in deep reinforcement learning, researchers developed sophisticated techniques to approximate state visitation frequencies. Early efforts, such as those by (5), demonstrated that deep predictive models could generate intrinsic exploration bonuses based on learned system dynamics in complex visual environments like Atari games. While not strictly count-based, this work highlighted the potential of neural networks to process high-dimensional observations and produce meaningful intrinsic signals, setting the stage for more direct approximations of novelty.

A pivotal breakthrough in scaling count-based exploration was the introduction of pseudo-counts and density models. (**?**) proposed a unified framework that generalizes count-based exploration by estimating state visitation frequencies using density models, thereby generating "pseudo-counts" for high-dimensional observations. Instead of exact state matching, this approach leverages the statistical likelihood of observing a state given past experiences. States that are less probable under the learned density model are considered more novel and receive higher intrinsic rewards. This effectively overcomes the limitations of exact state enumeration by providing a continuous and differentiable measure of novelty.

Building on this principle, various practical implementations emerged. (2) introduced `#Exploration`, a surprisingly effective yet simple method that maps high-dimensional states to hash codes. By counting the occurrences of these hash codes, the approach approximates state visitation frequencies, allowing for scalable pseudo-counting. This demonstrated that even a relatively crude approximation of state novelty, when combined with deep reinforcement learning, could yield near state-of-the-art performance on chal-

lenging benchmarks. However, the effectiveness of hash-based methods can be sensitive to the choice of hash function and the potential for hash collisions, which might conflate distinct states. Further refining the use of neural networks for density estimation, (**?**) explicitly employed neural density models to compute pseudo-counts. This provided a more principled and robust statistical approach to estimate state novelty in complex visual environments, as the density model can learn more meaningful representations of states and their relationships. The challenge with such methods lies in the computational complexity of training accurate density models in high-dimensional spaces and ensuring that the learned density truly reflects meaningful novelty rather than irrelevant stochasticity.

While count-based and density-based methods primarily focus on the frequency of state visitation, other intrinsic motivation techniques, such as curiosity-driven exploration, leverage prediction error as a proxy for novelty. Methods like the Intrinsic Curiosity Module (ICM) by (**?**) and Random Network Distillation (RND) by (**?**) reward agents for encountering states where their internal predictive models are inaccurate or for states that lead to unpredictable outcomes. These approaches offer an alternative perspective on novelty, focusing on the agent's learning progress or uncertainty about environmental dynamics rather than mere visitation frequency. Although distinct in their underlying signals, both paradigms share the common goal of generating intrinsic rewards to drive exploration in sparse-reward, high-dimensional settings, with density-based methods providing a statistical measure of "unvisitedness" and prediction-error methods focusing on "unpredictability."

In summary, count-based and density-based novelty methods have undergone a significant evolution, transforming from simple heuristics for discrete environments into sophisticated deep learning techniques capable of scaling to complex, high-dimensional state spaces. The transition from direct state counting to pseudo-counts derived from neural density models has been critical for enabling robust exploration in deep reinforcement learning. These techniques provide a practical and often effective way to incentivize broad state space coverage and discover new areas. Nevertheless, challenges persist, including the computational overhead of training accurate density models, the sensitivity

to state representation, and the difficulty of ensuring that the quantified novelty aligns with task-relevant exploration rather than being misled by uninformative stochasticity. Future research continues to refine these methods, often by integrating insights from both visitation statistics and predictive uncertainty, to develop more adaptive and robust novelty-seeking agents.

4.3 Prediction Error and Self-Supervised Curiosity

Effective exploration remains a cornerstone challenge in reinforcement learning (RL), particularly in environments characterized by sparse rewards and high-dimensional observations. To address this, intrinsic motivation methods have emerged, where agents generate their own reward signals to drive discovery. A prominent approach within this paradigm is self-supervised curiosity, which leverages the agent's ability to predict future states or features, using prediction error as an intrinsic reward to guide exploration. This strategy incentivizes agents to seek out situations where their internal models of the world are inaccurate, thereby driving learning about the environment's underlying dynamics.

The foundational concept of curiosity as a driver for learning can be traced back to early work by (**?**), which proposed that agents could be intrinsically motivated to explore by optimizing the predictability of their sensory inputs. This early idea laid the groundwork for defining curiosity as a measure of surprise or novelty. Expanding on this, (5) demonstrated that deep predictive models could be effectively used to assign exploration bonuses in complex domains like Atari games, by rewarding states where the agent's learned dynamics model exhibited high uncertainty. This represented an important step in scaling prediction-error-based curiosity to high-dimensional visual inputs, moving beyond simpler tabular settings.

A significant advancement in this direction was the Intrinsic Curiosity Module (ICM) proposed by (**?**). ICM defines curiosity as the error in predicting the consequence of an agent's own actions within a learned feature space. Specifically, it trains a self-supervised forward dynamics model to predict the next latent state given the current latent state and action. The magnitude of this prediction error then serves as the intrinsic

reward. This design incentivizes the agent to explore states where its internal model is inaccurate, thereby driving learning about the environment's dynamics. Crucially, by operating in a learned feature space rather than raw pixels, ICM made an initial attempt to mitigate the "noisy TV problem," where agents might be perpetually drawn to uncontrollable stochastic elements (like static on a TV screen) that generate high prediction error but offer no meaningful learning progress. This focus on learnable and controllable aspects of the environment represented a significant step towards scalable curiosity in high-dimensional visual environments.

Despite ICM's success, its reliance on learning an accurate forward dynamics model can still be problematic, particularly in highly stochastic environments. In such settings, genuine environmental noise or inherent unpredictability can lead to consistently high prediction errors, which are uninformative for learning and can still distract the agent, leading to inefficient exploration. This limitation highlights a critical distinction: prediction error can arise from either the agent's lack of knowledge (epistemic uncertainty) or from inherent environmental stochasticity (aleatoric uncertainty). ICM, by primarily measuring the error of a single forward model, struggles to differentiate between these two sources, potentially leading to spurious curiosity signals.

To address this challenge and provide more robust curiosity signals, alternative prediction-error-based approaches have emerged. One prominent direction involves leveraging ensembles of models to quantify epistemic uncertainty more explicitly. For instance, methods like Exploration via Distributional Ensemble (EDE) (139) encourage exploration of states with high epistemic uncertainty by using an ensemble of Q-value distributions. The disagreement or variance among the predictions of these ensemble members provides a more reliable signal of what the agent truly "doesn't know," rather than simply what is unpredictable due to noise. This ensemble-based approach offers a principled way to direct exploration towards areas where the agent's understanding of the environment is weakest, promoting more efficient knowledge acquisition. The concept of using prediction error as a curiosity signal is also versatile, extending to domains like Large Language Models, where signals such as perplexity over generated responses or variance of value estimates

from multi-head architectures can serve as intrinsic exploration bonuses (232). From an information-theoretic perspective, these methods align with the idea of maximizing information gain, where surprise (prediction error) and novelty drive the building of abstract dynamics models and transferable skills (34).

The versatility of these prediction-error-based curiosity mechanisms has led to their integration into various RL frameworks. For instance, (121) demonstrated how a simplified version of ICM could be effectively combined with off-policy RL methods, such as Deep Deterministic Policy Gradient (DDPG) and Hindsight Experience Replay (HER), significantly improving exploration efficiency and learning performance in robotic manipulation tasks with sparse rewards. Furthermore, the utility of curiosity-based intrinsic motivation extends to the challenging domain of offline reinforcement learning. (46) investigated how such curiosity-driven methods could be used to collect informative datasets in a task-agnostic manner, which could then be leveraged by offline RL algorithms, highlighting their role in generating high-quality data for subsequent learning.

While prediction error and self-supervised curiosity have proven highly effective in driving exploration by incentivizing agents to learn about their environment's dynamics, challenges remain. The primary limitation lies in distinguishing between genuine uncertainty that can be resolved through exploration and inherent environmental stochasticity that offers no meaningful learning progress. This distinction is crucial for designing intrinsic reward functions that consistently align with meaningful exploration, especially in complex, hierarchical tasks. The balance between intrinsic and extrinsic rewards, and the potential for agents to get stuck in "perpetual exploration" loops without making tangible task progress, are also critical considerations. Addressing the robustness of these intrinsic signals against uninformative stochasticity is a key area of ongoing research, motivating the development of more sophisticated methods that will be discussed in the subsequent section.

4.4 Robust Intrinsic Rewards: Addressing the Noisy TV Problem

The quest for effective exploration in reinforcement learning (RL) is profoundly challenged by environments offering sparse extrinsic rewards. While intrinsic motivation methods emerged as a promising avenue to generate internal curiosity signals and alleviate this sparsity, early prediction-error approaches frequently succumbed to the "noisy TV problem." This phenomenon describes scenarios where agents are perpetually drawn to unpredictable yet uninformative stochastic elements within the environment, such as random pixel noise on a screen or flickering lights. These elements generate consistently high prediction errors, leading to spurious curiosity that distracts the agent from truly novel and learnable aspects of the environment, thereby hindering focused exploration.

Initial efforts, such as those by (5), explored incentivizing exploration through bonuses derived from concurrently learned deep predictive models of system dynamics. This aimed to reward agents for encountering states that challenged their current environmental understanding. A prominent example of this paradigm is the Intrinsic Curiosity Module (ICM) (**?**). ICM trains a forward dynamics model to predict the next state's features given the current state's features and the executed action. The magnitude of this prediction error then serves as an intrinsic reward, encouraging the agent to visit states where its internal model is inaccurate. While ICM provided a scalable solution for high-dimensional visual inputs by operating in a learned feature space, its reliance on predicting *future states* made it inherently susceptible to the noisy TV problem. In environments with uninformative stochasticity, the forward dynamics model would consistently fail to predict these truly random, irreducible changes, resulting in persistently high prediction errors. This spurious curiosity would then cause the agent to repeatedly visit these uninformative areas, diverting computational resources and hindering progress towards task-relevant exploration.

To overcome the critical limitations posed by the noisy TV problem, (**?**) introduced Random Network Distillation (RND), a pivotal innovation in robust intrinsic reward generation. RND fundamentally redefines novelty detection by decoupling the intrinsic reward from the learnability of the environment's true dynamics. Instead of predicting

future states, RND employs two neural networks: a fixed, randomly initialized target network (f) and a predictor network ($\hat{f}$). Both networks receive the current state s_t as input. The predictor network is trained to predict the output of the target network, i.e., $\hat{f}(s\_t) \approx f(s\_t)$. The intrinsic reward is then defined as the mean squared error between the outputs of these two networks: $r\_i = ||\hat{f}(s\_t) - f(s\_t)||^2$.

The key insight of RND lies in its design: for any given state s_t, even one containing uninformative stochasticity (like a noisy TV screen), the randomly initialized target network $f(s\_t)$ produces a *fixed and deterministic* output embedding. The predictor network $\hat{f}(s\_t)$ is then trained to learn this fixed mapping. If an agent repeatedly visits a state s_t, the predictor $\hat{f}$ will eventually learn to accurately map s_t to $f(s\_t)$, causing the prediction error and thus the intrinsic reward to *decay*. This mechanism effectively filters out irrelevant stochasticity because the reward is based on the *novelty of the state itself* (i.e., how well the predictor has learned to map that specific state to its fixed target), rather than the unpredictability of state transitions. Consequently, even a noisy TV state, despite its visual randomness, yields a fixed target output from f. Once the agent has sufficiently explored this state, $\hat{f}$ learns the mapping, and the curiosity bonus diminishes, preventing perpetual attraction. This contrasts sharply with ICM, where the prediction error for a truly stochastic transition remains irreducible, leading to persistent curiosity.

RND's robust and simpler mechanism for novelty detection proved highly effective, leading to significantly more efficient and directed exploration in complex visual domains, such as Atari games, where it substantially outperformed prior methods on hard exploration tasks (**?**). Its success underscored the importance of designing intrinsic reward signals that are resilient to environmental noise and focus on aspects of the environment truly conducive to learning. Subsequent works, such as (121), further explored simplified variants of curiosity modules, demonstrating how architectural or methodological simplifications could enhance the practical utility and integration of prediction-error based intrinsic motivation, particularly for off-policy reinforcement learning methods.

While RND provided a powerful solution to the noisy TV problem, it is not the sole robust intrinsic motivation strategy. Other approaches also aim to mitigate the effects

of uninformative stochasticity or enhance exploration in complementary ways. For instance, methods based on model disagreement or ensembles, which reward exploration in states where multiple learned dynamics models disagree, can implicitly discount uncontrollable noise by focusing on areas where the agent's *learnable* understanding is inconsistent. Information-theoretic perspectives, as surveyed by (34), often emphasize maximizing *useful* information gain, which aligns with RND's implicit discounting of unlearnable noise. Furthermore, diversity-driven exploration strategies, which incentivize agents to visit states that are distinct from previously encountered ones (20), or Random Latent Exploration (RLE), which encourages agents to pursue randomly sampled goals in a latent space (187), offer alternative mechanisms for robust and deep exploration without relying solely on prediction error. Deep Curiosity Search (DeepCS) (76) also introduced the concept of "intra-life novelty," rewarding exploration within a single episode, which can complement RND's "across-training novelty" by encouraging immediate discovery.

Despite these significant advancements, challenges persist. While RND effectively addresses the noisy TV problem by focusing on learnable novelty, the intrinsic reward signal can still be somewhat undirected, potentially leading to exploration of areas that are novel but not necessarily relevant to the extrinsic task. Future research continues to explore how to imbue intrinsic rewards with more task-relevant directionality, perhaps through goal-conditioned or hierarchical approaches, or how to combine them with other exploration strategies to achieve even more efficient and purposeful exploration in increasingly complex and open-ended environments. The integration of RND into advanced agents like Agent57 (**?**) further demonstrates its lasting impact as a foundational component for achieving state-of-the-art performance in diverse and challenging domains.

5 Advanced and Adaptive Exploration Strategies

5.1 Hierarchical Reinforcement Learning for Exploration

Effective exploration is a critical bottleneck in reinforcement learning (RL), particularly in complex environments characterized by vast state-action spaces, sparse rewards, and long

task horizons. Hierarchical Reinforcement Learning (HRL) offers a powerful and principled framework to address these challenges by decomposing large, intractable problems into a hierarchy of more manageable sub-problems. This decomposition allows agents to learn and explore at different levels of temporal abstraction, significantly enhancing exploration efficiency and robustness.

The foundational concept underpinning HRL for exploration is the "Options Framework" introduced by (?). Building upon earlier notions of skill chaining by (?), the Options Framework formalizes "options" as temporally extended actions, or sub-policies, that execute for multiple time steps. An option consists of an initiation set (states where it can be taken), a policy (how to act while the option is active), and a termination condition (when the option ends). This framework allows a high-level policy to choose among options, while a low-level policy executes the primitive actions within a chosen option. This mechanism fundamentally aids exploration by enabling agents to traverse large portions of the state space more efficiently than with primitive actions alone. Instead of exploring individual steps, the agent explores sequences of actions (options), effectively reducing the effective search space and allowing for more directed movement towards relevant subgoals or novel regions. For instance, an agent might learn an "open door" option, which, once selected, reliably executes the necessary primitive actions to open a door, allowing the high-level policy to explore the consequences of being on the other side of the door, rather than stumbling upon the correct sequence of door-opening actions by chance.

In the era of deep reinforcement learning, various architectures have been proposed to implement hierarchical control and leverage its benefits for exploration. FeUdal Networks (FuN) (?) introduced a two-level hierarchy with a "Manager" module that sets goals in a latent space and a "Worker" module that executes primitive actions to achieve those goals. The Manager explores the space of goals, while the Worker explores the primitive action space conditioned on the current goal. This explicit goal-setting mechanism intrinsically guides exploration towards achieving meaningful sub-objectives. Similarly, Hierarchical Reinforcement Learning with Off-policy Correction (HIRO) (?) enables efficient learning of both high-level and low-level policies by correcting for off-policy data,

allowing the high-level policy to explore by setting goals in the state space, and the low-level policy to learn how to reach those goals. Hierarchical Actor-Critic (HAC) (**?**) further refines this by using multiple layers of actor-critic agents, where higher levels set goals for lower levels, and a "hindsight experience replay" mechanism allows agents to learn from failed attempts to reach goals, thereby improving exploration by making better use of suboptimal trajectories. These deep HRL methods demonstrate how learning goal-conditioned policies at different levels of abstraction can significantly accelerate exploration in complex, high-dimensional environments.

A critical aspect of HRL for exploration is the autonomous discovery of useful options or skills, often driven by intrinsic motivation. Rather than manually defining options, agents can learn them through self-supervision. Methods like Diversity is All You Need (DIAYN) (**?**) and Variational Option Discovery (VALOR) (**?**) learn a diverse set of skills by maximizing the mutual information between the skill executed and the resulting state trajectory, effectively rewarding the agent for discovering distinct behaviors. These learned skills then serve as valuable options for a higher-level policy, enabling more structured and efficient exploration. The survey by (34) highlights how information-theoretic intrinsic motivation, particularly novelty and surprise, can assist in building a hierarchy of transferable skills, making the exploration process more robust. Furthermore, (234) emphasizes unsupervised skill acquisition as a key advancement for enhancing scalability in open-ended environments, where HRL provides a natural framework for organizing these learned behaviors.

HRL also facilitates temporally coordinated and goal-conditioned exploration. (109) proposed Generative Planning Method (GPM), which generates multi-step action plans, effectively acting as temporally extended options. These plans guide exploration towards high-value regions more consistently than single-step perturbations, and the plan generator can adapt to the task, further benefiting future explorations. This aligns with HRL's ability to create intentional action sequences for reaching specific subgoals. Similarly, Random Latent Exploration (RLE) (187), while not strictly HRL, encourages exploration by pursuing randomly sampled goals in a latent space. This goal-conditioned exploration

paradigm is inherently compatible with HRL, where the high-level policy can sample latent goals for the low-level policy to achieve, fostering diverse and deep exploration.

The utility of hierarchical exploration extends significantly to multi-agent systems, where coordination and efficient search are paramount. (4) designed a cooperative exploration strategy for multiple mobile robots using a hierarchical control architecture, where a high-level decision-making layer coordinates exploration to minimize redundancy, and a low-level layer handles target tracking and collision avoidance. This demonstrates how HRL can structure complex multi-robot behaviors for efficient, coordinated exploration. More recently, (57) tackled cooperative visual exploration for multiple agents with a Multi-agent Spatial Planner (MSP) leveraging a transformer-based architecture with hierarchical spatial self-attentions, enabling agents to capture spatial relations and plan cooperatively based on visual signals. (176) further advances multi-agent exploration with "Imagine, Initialize, and Explore" (IIE), which uses a transformer to imagine critical states and then initializes agents at these states for targeted exploration. This approach, while not explicitly called HRL, embodies hierarchical decomposition by first identifying high-level critical states (subgoals) and then focusing low-level exploration from those points, enhancing the discovery of successful joint action sequences in long-horizon tasks.

In summary, HRL provides a powerful framework for managing the complexity of exploration in large state-action spaces. By enabling agents to learn and compose temporally extended actions (options) or to decompose complex tasks into sub-problems, HRL significantly reduces the dimensionality of the exploration problem. This structured approach allows agents to focus on achieving meaningful subgoals, leading to more directed and sample-efficient discovery of optimal policies in long-horizon tasks. Despite these advancements, challenges persist, particularly in the autonomous discovery of optimal and diverse skill sets, the robust learning of high-level policies that effectively coordinate lower-level skills, and ensuring seamless communication and transfer of information across hierarchical levels. Future research will likely focus on developing more adaptive and autonomous methods for skill acquisition and composition, further integrating HRL with robust intrinsic motivation and meta-learning to create agents capable of truly open-

ended and efficient exploration.

5.2 Learning Exploration Policies (Meta-Exploration)

A pivotal advancement in reinforcement learning (RL) exploration shifts the paradigm from relying on hand-crafted heuristics or fixed intrinsic reward functions to enabling agents to autonomously learn their own exploration strategies. This sophisticated approach, termed meta-exploration, involves an outer-loop optimization process that trains a meta-controller or a recurrent policy to generate adaptive exploration behaviors, aiming to maximize long-term returns across a distribution of tasks or episodes. By learning "how to explore," these methods can dynamically adjust the exploration-exploitation trade-off, leading to more efficient, task-relevant discovery and robust performance, particularly in novel and complex environments (**? ?**). This represents a significant stride towards autonomous and intelligent exploration, where agents generalize their exploration capabilities rather than relearning them from scratch for each new task.

The foundational concept of meta-exploration was significantly advanced by works demonstrating that recurrent neural networks (RNNs) can serve as meta-learners. (**?**) introduced RL^2 (Reinforcement Learning to Reinforce Learn), a seminal framework where an RNN-based agent is trained to solve a distribution of tasks. The RNN's hidden state effectively encodes task-specific information and past experience, allowing it to learn an exploration strategy that adapts within a single episode and across multiple episodes of a new, unseen task. This enables the agent to exhibit rapid adaptation and efficient exploration behaviors, such as directed search or uncertainty-driven probing, without explicit hand-engineered exploration bonuses. Similarly, (**?**) explored the idea of "Learning to Reinforce Learn," where an RNN acts as a meta-learner to discover an entire RL algorithm, including its exploration component, by processing sequences of observations, actions, and rewards. These approaches highlight the power of recurrent architectures to implicitly capture and execute sophisticated exploration policies that generalize across related tasks.

Building on these foundations, subsequent research has focused on learning more struc-

tured and informed exploration strategies. (6) proposed Model Agnostic Exploration with Structured Noise (MAESN), a gradient-based meta-learning algorithm that learns exploration strategies from prior experience. MAESN leverages prior tasks to initialize a policy and acquire a latent exploration space, which injects structured stochasticity into the policy. This allows for exploration strategies that are informed by previous knowledge, moving beyond simple action-space noise and proving more effective than task-agnostic methods. This work underscores the benefit of meta-learning not just the policy, but the *mechanism* of exploration itself, enabling more targeted and efficient discovery.

Meta-learning has also been applied to the generation and refinement of intrinsic motivation signals for exploration. (80) introduced MetaCURE (Meta Reinforcement Learning with Empowerment-Driven Exploration), which explicitly models an exploration policy learning problem separate from the exploitation policy. MetaCURE employs a novel empowerment-driven exploration objective that aims to maximize information gain for task identification, deriving a corresponding intrinsic reward. By learning separate, context-aware exploration and exploitation policies and sharing task inference knowledge, MetaCURE significantly enhances exploration efficiency in sparse-reward meta-RL tasks. This demonstrates how meta-learning can discover effective intrinsic reward functions that guide exploration towards truly informative experiences, addressing a key challenge in intrinsic motivation.

Furthermore, meta-exploration has been integrated with other advanced RL paradigms. (162) presented MAMBA, a model-based approach to meta-RL that leverages world models for efficient exploration. By learning an internal model of the environment, MAMBA can plan and explore more effectively, leading to greater return and significantly improved sample efficiency (up to 15x) compared to existing meta-RL algorithms, especially in higher-dimensional domains. This highlights the synergy between learning environmental models and meta-learning exploration strategies. Complementing this, (167) introduced Learned Optimization for Plasticity, Exploration and Non-stationarity (OPEN), which meta-learns an update rule (an optimizer) whose input features and output structure are informed by solutions to common RL difficulties, including exploration. OPEN's pa-

rameterization is flexible enough to use stochasticity for exploration, demonstrating that meta-learning can discover effective policy update mechanisms that inherently promote efficient exploration.

It is crucial to distinguish meta-exploration from merely adaptive exploration, where exploration parameters are tuned within a single learning process. While methods that dynamically adjust exploration probabilities based on metrics like information entropy (118) or ensemble learning for balancing exploration-exploitation ratios (159) are valuable, they typically do not involve an outer-loop meta-training process to learn a generalizable exploration *strategy* across tasks. Such adaptive approaches are better categorized under integrated and adaptive exploration frameworks (Section 5.3), which focus on dynamic parameter adjustment rather than learning the exploration algorithm itself.

In conclusion, the shift towards learning exploration policies through meta-exploration represents a profound step towards truly autonomous and intelligent agents. These approaches, ranging from recurrent policies learning exploration behaviors across episodes to meta-learning structured exploration noise, intrinsic motivation signals, or even entire optimization rules, empower agents to generalize their exploration capabilities and achieve robust performance across diverse problem settings. However, significant challenges persist, including the computational expense of meta-training, the difficulty of defining appropriate and diverse task distributions for meta-learning, ensuring the learned exploration strategies generalize to truly novel and out-of-distribution tasks, and designing effective meta-objectives that capture desirable exploration properties. Future research will likely focus on developing more robust, sample-efficient, and generalizable meta-learning algorithms that can discover truly novel and effective exploration strategies across a wide spectrum of tasks, pushing the boundaries of autonomous discovery.

5.3 Integrated and Adaptive Exploration Frameworks

Effective exploration in complex, high-dimensional environments often transcends the capabilities of a single, static strategy, necessitating frameworks that dynamically combine multiple exploration techniques and adaptively select or blend them based on the current

task, state, or learning progress. Moving beyond a 'one-size-fits-all' approach, these integrated frameworks aim to create highly versatile and robust agents capable of tackling a wide spectrum of exploration challenges, representing a key direction for developing general-purpose reinforcement learning (RL) agents. Robust intrinsic motivation, as discussed in Section 4, frequently serves as a foundational component within these integrated systems, providing internal reward signals (e.g., RND-style novelty (**?**) or episodic novelty (**?**)) that are then managed or orchestrated by higher-level adaptive mechanisms.

A prominent approach to achieving adaptive exploration involves the use of meta-controllers that explicitly orchestrate a portfolio of exploration-exploitation policies. The seminal work of (**?**) on Agent57 exemplifies this paradigm, achieving state-of-the-art performance across diverse Atari games. Agent57 integrates multiple intrinsic motivation signals, including Random Network Distillation (RND) for life-long novelty and value-discrepancy-based novelty, with an adaptive exploration strategy managed by a meta-controller. This meta-controller dynamically selects from a range of exploration-exploitation policies, allowing the agent to adjust its behavior on the fly to suit the specific demands of each game and phase of learning. The strength of this approach lies in its ability to explicitly learn *how* to explore by selecting appropriate behaviors from a predefined set, offering significant flexibility. However, a limitation is its reliance on a hand-designed portfolio of base policies and the complexity of meta-learning the controller, which can be computationally intensive and may struggle if the optimal strategy is not represented within the initial portfolio.

Beyond explicit meta-controllers, other integrated frameworks leverage ensemble methods or decoupled architectures to achieve robust and adaptive exploration. Ensemble-based approaches, for instance, integrate multiple policies or value functions to capture uncertainty or promote diversity. (139) introduced Exploration via Distributional Ensemble (EDE), a method that encourages exploration of states with high epistemic uncertainty through an ensemble of Q-value distributions. This integration of multiple distributional estimates allows for a more nuanced understanding of uncertainty, leading to improved generalization in unseen environments. Similarly, (59) proposed Ensemble Proximal Policy

Optimization (EPPO), which learns ensemble policies in an end-to-end manner, combining individual policies and the ensemble organically. EPPO adopts a diversity enhancement regularization over the policy space, which theoretically increases exploration efficacy and promotes generalization. In a different vein, (98) proposed Decoupled Exploration and Exploitation Policies (DEEP), which structurally separates the task policy from the exploration policy. This decoupling allows for directed exploration to be highly effective for sample-efficient continuous control without incurring performance penalties in densely-rewarding environments. In contrast to Agent57's explicit switching mechanism, these ensemble and decoupled architectures offer a more implicit form of adaptation, either through aggregation of diverse perspectives or a clear separation of learning objectives, providing robustness but potentially less fine-grained, task-specific adaptation.

Further advancements in adaptive exploration leverage probabilistic, Bayesian, and reactive mechanisms to guide behavior based on uncertainty or environmental shifts. Bayesian approaches provide a principled framework for managing uncertainty, which can be directly used to drive exploration. (90) introduced a model-based lifelong RL approach that estimates a hierarchical Bayesian posterior, which, combined with a sample-based Bayesian exploration procedure, adaptively increases sample efficiency across related tasks. Extending this, (204) explored Bayesian exploration with Implicit Posteriori Parameter Distribution Optimization (IPPDO), modeling parameter uncertainty with an implicit distribution approximated by generative models, offering greater flexibility and improved sample efficiency. In the realm of randomized exploration, (148) and (189) developed scalable Thompson sampling strategies using Langevin Monte Carlo and approximate sampling, respectively. These methods directly sample the Q-function from its posterior distribution, providing provably efficient and adaptive exploration by inherently balancing uncertainty and reward. For dynamic environments, (92) proposed Reactive Exploration, designed to track and react to continual domain shifts in lifelong reinforcement learning, demonstrating adaptive policy updates in non-stationary settings. These probabilistic and reactive methods offer strong theoretical grounding for continuous adaptation but often face computational challenges in high-dimensional settings, requiring efficient

approximation techniques.

Emerging trends also point towards exploration as an emergent property from learned optimizers or large pre-trained models. (167) introduced Learned Optimization for Plasticity, Exploration and Non-stationarity (OPEN), a meta-learned update rule whose input features and output structure are informed by solutions to RL difficulties, including the ability to use stochasticity for exploration. This represents a shift towards learning the optimization process itself, which implicitly includes adaptive exploration. In the context of large foundation models, (128) demonstrated that Decision-Pretrained Transformers (DPT) can exhibit emergent online exploration capabilities in-context, without explicit training for it. Building on this, (158) proposed In-context Exploration-Exploitation (ICEE), which optimizes the efficiency of in-context policy learning by performing exploration-exploitation trade-offs at inference time within a Transformer model. These approaches suggest a future where exploration is less explicitly engineered and more implicitly learned or emergent, potentially leading to highly generalizable agents, but raise questions about control, interpretability, and the sample efficiency required for pre-training.

In conclusion, integrated and adaptive exploration frameworks represent a significant leap towards developing general-purpose RL agents. By combining robust intrinsic motivation signals, explicit meta-controllers, ensemble and decoupled architectures, principled Bayesian and probabilistic methods, and leveraging emergent capabilities from learned optimizers and large models, these frameworks move beyond static heuristics to dynamically adjust their exploration behaviors. While significant progress has been made, future directions include developing more theoretically grounded guarantees for these complex adaptive systems, enhancing their computational efficiency and scalability, and ensuring robust generalization to truly novel and open-ended environments where the optimal exploration strategy might not be easily predefined or learned from limited prior experience. A key unresolved challenge is the automated discovery of an optimal portfolio of exploration strategies for meta-controllers, as current approaches often rely on hand-designed components.

5.4 Population-Based and Evolutionary Exploration

The inherent challenge of the exploration-exploitation dilemma in Reinforcement Learning (RL) often leads single agents to converge prematurely to sub-optimal policies, particularly in environments characterized by sparse rewards or complex, multi-modal reward landscapes. To overcome these limitations, a distinct paradigm has emerged that leverages populations of agents or evolutionary algorithms to foster broader, more robust exploration and facilitate global search. These approaches represent a meta-level solution, structuring the learning system itself for enhanced discovery.

Early precursors to population-based exploration can be found in Neuroevolution, where neural network architectures and weights are optimized using evolutionary algorithms. Methods like NEAT (NeuroEvolution of Augmenting Topologies) (**?**) demonstrated the power of evolving diverse populations of networks to solve complex control tasks, implicitly performing exploration by searching a vast hypothesis space. More recently, Evolution Strategies (ES) have gained prominence as a scalable black-box optimization technique for deep RL, capable of training deep neural networks efficiently due to their high parallelizability (**?**). However, standard ES can struggle with sparse or deceptive reward landscapes, necessitating directed exploration. To address this, (9) introduced methods like Novelty Search with Evolution Strategies (NS-ES) and Quality Diversity (QD) algorithms hybridized with ES. These approaches maintain a population of novelty-seeking agents, rewarding exploration of novel behaviors rather than just high performance, thereby enabling ES to avoid local optima and achieve higher performance on challenging deep RL tasks like Atari games and simulated robot locomotion.

A significant advancement in population-based methods for deep RL is Population Based Training (PBT) (**?**). PBT concurrently trains a population of agents, each with its own set of hyperparameters and model weights. Unlike traditional grid search or random search, PBT dynamically adapts hyperparameters during training by periodically evaluating agents, exploiting well-performing ones (copying their weights and hyperparameters) and exploring new hyperparameter configurations for underperforming ones. This asynchronous "exploit-and-explore" strategy allows PBT to discover robust hyper-

parameter schedules and model weights simultaneously, leading to faster training and improved final performance across diverse tasks, including complex deep RL benchmarks. PBT's strength lies in its ability to adaptively tune both learning processes and agent policies, making it highly effective for complex, high-dimensional problems.

Building on the strengths of both evolutionary algorithms and gradient-based RL, Evolutionary Reinforcement Learning (ERL) frameworks have emerged as a powerful hybrid paradigm. These methods typically combine the global search capabilities of evolutionary algorithms with the local optimization efficiency of gradient-based RL. Early ERL approaches, such as those by (**?**) and CEM-RL (**?**), often involve evolving a population of actor networks, while a shared critic network (trained via gradient descent) provides value estimates to guide both the evolutionary process and individual actor updates. This hybrid approach aims to leverage the exploration benefits of evolution (e.g., escaping local optima, maintaining diversity) and the sample efficiency of RL. For instance, (223) proposed a modified ERL method for multi-agent region protection, amalgamating Differential Evolution (DE) for diverse sample exploration and overcoming sparse rewards with Multi-Agent Deep Deterministic Policy Gradient (MADDPG) for training defenders and expediting DE convergence.

A more recent advancement in this line is the Two-Stage Evolutionary Reinforcement Learning (TERL) framework proposed by (219). TERL addresses a key limitation of prior ERL methods, which often evolve only actor networks, thereby constraining exploration if a single critic network falls into local optima. Instead, TERL maintains and optimizes a population of *complete RL agents*, each comprising both an actor and a critic network. This design enables more independent and diverse exploration by each individual, mitigating the risk of premature convergence to suboptimal policies dictated by a flawed shared critic. The TERL framework operates through a novel two-stage learning process: an initial "Exploration Stage" where all individuals learn independently, optimized by a hybrid approach combining gradient-based RL updates with meta-optimization techniques like Particle Swarm Optimization (PSO). This stage emphasizes diversification, with agents sharing information efficiently through a common replay buffer, which helps propagate

beneficial experiences across the population. Following this, the "Exploitation Stage" focuses on refining the best-performing individual from the population through concentrated RL-based updates, while the remaining individuals continue to undergo PSO to further diversify the replay buffer. This dynamic allocation of computational resources and tailored optimization strategies across stages allows TERL to effectively balance the exploration-exploitation dilemma.

Despite their advantages, population-based and evolutionary methods introduce their own set of challenges. Computational cost is a primary concern, as maintaining and training multiple agents or evolving large populations can be resource-intensive, although parallelization strategies (like those in ES and PBT) mitigate this. Furthermore, the integration of diverse data from population optimization into off-policy RL algorithms, particularly through shared replay buffers, can introduce instability and even degrade performance, as highlighted by (185). This issue arises because population data, while diverse, might not align with the on-policy distribution expected by some RL algorithms, leading to an "overlooked error." To remedy this, (185) proposed a double replay buffer design to provide more on-policy data, demonstrating the need for careful architectural considerations when combining these paradigms. The choice between PBT's hyperparameter evolution and ERL's policy evolution also presents a trade-off: PBT excels at finding robust training configurations, while ERL directly optimizes policy parameters, often leading to more direct policy improvement.

In conclusion, population-based and evolutionary exploration methods offer a compelling meta-level solution to the challenges of exploration in complex RL environments. By evolving populations of complete RL agents, dynamically adapting hyperparameters, or employing hybrid optimization strategies, these approaches enable more diverse learning trajectories and a more robust search for optimal policies, moving beyond the limitations of single-agent exploration heuristics. Future research could explore more sophisticated mechanisms for inter-agent information sharing, investigate adaptive intrinsic motivation signals within these population-based frameworks, or extend these concepts to multi-task and open-ended learning scenarios, further enhancing the adaptability and

generalization capabilities of RL agents.

6 Specialized Contexts and Applications of Exploration

6.1 Exploration in Offline Reinforcement Learning

In offline Reinforcement Learning (RL), the agent is tasked with learning an optimal policy solely from a fixed, pre-collected dataset, fundamentally precluding any further active interaction with the environment. This paradigm shift introduces unique challenges for exploration, as traditional active exploration strategies, which involve generating new experiences, are inherently impossible. Instead, the core problem transforms into managing *distributional shift* and preventing the policy from querying actions that are out-of-distribution (OOD) relative to the dataset. The focus shifts from active exploration to ensuring 'conservative learning' within the boundaries of the existing data distribution, aiming to identify reliable regions for policy improvement while rigorously avoiding OOD actions that could lead to unreliable value estimates or unsafe behaviors due to extrapolation errors. This conservative approach is paramount for real-world applications where data collection is costly, risky, or simply not feasible during the learning process.

The foundational approaches to offline RL primarily address the distributional shift problem through two main mechanisms: policy constraints and value function pessimism. Seminal works like Batch-Constrained Q-Learning (BCQ) (**?**) introduced explicit policy constraints, regularizing the learned policy to stay close to the behavior policy that generated the dataset. This is typically achieved by adding a regularization term (e.g., KL-divergence) to the policy objective, ensuring that the agent does not venture into unobserved state-action pairs. Similarly, Behavior Regularized Actor Critic (BRAC) (**?**) further explored various forms of behavior regularization to mitigate the distributional shift. While effective in keeping the policy close to the data, these methods can sometimes limit the discovery of truly optimal policies if the behavior policy was suboptimal.

A complementary and highly influential approach is Conservative Q-Learning (CQL) (**?**). CQL tackles the distributional shift by explicitly enforcing pessimism in the value

function estimation. It achieves this by adding a penalty to the Q-values of OOD actions, ensuring that the learned Q-function provides a lower bound on the true Q-values. This prevents overestimation of action values for unseen actions, which is a common failure mode in offline RL. While BCQ primarily constrains the policy directly, CQL intervenes at the value function level, indirectly shaping the policy by making OOD actions less attractive. This distinction highlights different points of intervention for ensuring conservatism.

Building upon these foundations, subsequent methods have leveraged uncertainty estimation to guide conservative learning more explicitly. The Uncertainty Weighted Actor-Critic (UWAC) (8) explicitly incorporates uncertainty treatment by detecting OOD state-action pairs and down-weighting their contribution in the training objectives. Utilizing a practical dropout-based uncertainty estimation, UWAC laid groundwork for robust learning by mitigating the impact of unreliable data points. Similarly, (36) conceptualized offline RL as "anti-exploration," proposing to subtract a prediction-based exploration bonus from the reward. This innovative approach encourages the policy to remain within the support of the dataset by penalizing actions whose consequences cannot be reliably predicted, effectively extending pessimism-based offline RL methods to deep learning settings. Further, (140) introduced an Entropy-regularized Diffusion Policy with Q-Ensembles, where robust policy improvement is achieved by learning the lower confidence bound of Q-ensembles. This implicitly accounts for uncertainty in value estimates, mitigating the impact of inaccurate value functions from OOD data, while an entropy regularizer improves "exploration" within the offline dataset by encouraging diverse actions within the observed distribution.

Simpler yet effective mechanisms have also emerged to ensure in-sample learning. (181) proposed Improving Offline Reinforcement Learning with in-Sample Advantage Regularization (ISAR), which adapts offline RL to robotic manipulation with minimal changes. ISAR learns the state-value function exclusively from dataset samples, then calculates the advantage function based on this in-sample estimation and adds a behavior cloning regularization term. This method effectively mitigates the impact of unseen actions without

introducing complex hyperparameters, offering a straightforward approach to conservative learning that implicitly handles OOD issues.

A significant challenge, particularly in model-based offline RL (MBRL), is addressing biased exploration during the synthetic trajectory generation phase. Standard maximum entropy exploration mechanisms, often adopted from online RL, can lead to skewed data distributions and impaired performance when applied to learned dynamics models. To tackle this, (229) introduced OCEAN-MBRL (Offline Conservative ExplorAtioN for Model-Based Offline Reinforcement Learning), a novel plug-in rollout approach. OCEAN explicitly decouples exploration from exploitation, introducing a principled, conservative exploration strategy guided by an ensemble of dynamics models for uncertainty estimation. It employs three key constraints: a state evaluation constraint to explore only in low-uncertainty regions, an exploration range constraint to select conservative actions, and a trajectory truncation constraint to limit rollouts in high-uncertainty areas. This comprehensive approach significantly enhances the stability and performance of existing MBRL algorithms by ensuring that exploration within the learned model remains reliable and does not generate new, potentially unsafe, out-of-distribution experiences.

The transition from offline learning to online fine-tuning presents another critical juncture for conservative exploration. While initial conservatism is vital for stable offline learning, a purely pessimistic policy might fail to discover better actions during online interaction. The Simple Unified uNcertainty-Guided (SUNG) framework for offline-to-online RL (199) quantifies uncertainty via a VAE-based state-action visitation density estimator. SUNG's adaptive exploitation method applies conservative offline RL objectives to high-uncertainty samples and standard online RL objectives to low-uncertainty samples, demonstrating a sophisticated use of uncertainty to dynamically adjust the degree of conservatism. Building on this, (149) proposed Optimistic Exploration and Meta Adaptation (OEMA) for sample-efficient offline-to-online RL. OEMA employs an optimistic exploration strategy, adhering to the principle of optimism in the face of uncertainty, allowing agents to sufficiently explore while reducing distributional shift through meta-learning. Providing a theoretical underpinning for such adaptive strategies, (225) showed

that Bayesian design principles are crucial for offline-to-online fine-tuning, suggesting that a probability-matching agent, rather than purely optimistic or pessimistic ones, can avoid sudden performance drops while being guaranteed to find the optimal policy.

The literature on exploration in offline RL has thus evolved from foundational methods that strictly constrain policies or penalize OOD value estimates to sophisticated, explicit conservative exploration strategies, particularly in model-based settings and during the offline-to-online transition. While significant progress has been made in ensuring learning remains robust and effective without active interaction, a persistent challenge lies in the accuracy, reliability, and computational efficiency of uncertainty estimation itself. Future research directions could focus on developing more robust and scalable uncertainty quantification methods, as well as adaptive mechanisms that dynamically adjust conservative exploration constraints based on the evolving confidence in the learned policy and model, effectively balancing the need for safety with the potential for performance improvement.

6.2 Expert-Guided and Demonstration-Based Exploration

Efficient exploration remains a formidable challenge in reinforcement learning (RL), particularly in complex, high-dimensional, and sparse-reward environments. To mitigate this, a significant body of research focuses on leveraging expert knowledge or demonstrations to guide and accelerate the exploration process, thereby improving sample efficiency and policy robustness. This paradigm, often termed Reinforcement Learning from Demonstrations (RLfD), seeks to bridge the gap between purely autonomous trial-and-error and the wealth of human or simulated expertise.

Early efforts established the foundational utility of demonstrations in overcoming exploration hurdles. (1) demonstrated that even a small set of expert demonstrations, when integrated with RL algorithms like Deep Deterministic Policy Gradients (DDPG) and Hindsight Experience Replay (HER), could provide an order of magnitude speedup in learning complex, continuous control robotics tasks with sparse rewards. These methods often relied on static, *offline* datasets of expert trajectories, which provided initial guidance but presented inherent limitations. Building upon this, (11) proposed Jump-Start

Reinforcement Learning (JSRL), which uses a "guide-policy" derived from offline data or demonstrations to form a curriculum of starting states for an "exploration-policy," significantly improving sample complexity, especially in data-scarce regimes. Similarly, (21) highlighted key ingredients for accelerating visual model-based RL with demonstrations, including policy pretraining, targeted exploration, and oversampling demonstration data, leading to substantial performance gains in sparse reward tasks.

Beyond direct trajectory imitation, expert knowledge can also be integrated through symbolic rules or domain-specific insights. For instance, (62) introduced Rule-Aware Reinforcement Learning (RARL) for knowledge graph reasoning, injecting high-quality symbolic rules into the model's reasoning process to alleviate sparse rewards and prevent spurious paths. (209) proposed using "state-action permissibility" knowledge to guide exploration, drastically speeding up deep RL training by identifying and avoiding impermissible actions. In a similar vein, (107) combined human knowledge-based rule bases with imitation learning pre-training (ILDN) and safe RL to enhance efficiency and generalization in large-scale adversarial scenarios. Furthermore, theoretical frameworks have explored how exploration itself can be framed as a utility to be optimized, which demonstrations can implicitly help achieve (15; 190).

Despite the benefits of offline demonstrations, a persistent challenge is the "distribution gap" (168). This occurs when the agent's policy deviates from the expert's, leading it into states not covered by the static demonstration dataset, thereby hindering generalization and robustness. To address this, research has shifted towards more dynamic and interactive integration of expert knowledge. (125) showed that with minimal modifications, existing off-policy RL algorithms could effectively leverage offline data (including demonstrations) for online learning, achieving significant performance improvements. This offline-to-online fine-tuning paradigm is crucial for real-world applications (166; 225), where policies pre-trained on demonstrations need to adapt to novel online experiences. For example, (138) presented VLA-RL, an approach that uses online reinforcement learning to improve pretrained Vision-Language-Action (VLA) models, specifically addressing out-of-distribution failures that arise from limited offline data in robotic manipulation.

Even leveraging "negative demonstrations" or failed experiences can guide exploration by showing what *not* to do, as demonstrated by (236) in sparse reward environments.

A significant advancement in bridging the distribution gap and enhancing exploration efficiency comes from dynamically interacting with experts. (240) introduced EARLY, an active RL from demonstrations algorithm where the agent intelligently queries for episodic demonstrations based on its trajectory-level uncertainty. This approach makes expert guidance more targeted and resource-efficient by requesting help only when needed. The pinnacle of this dynamic interaction is exemplified by (168) with RLfOLD (Reinforcement Learning from Online Demonstrations) for urban autonomous driving. RLfOLD introduces the novel concept of *online demonstrations*, which are dynamically collected from a simulator's privileged information during the agent's active exploration. These demonstrations are seamlessly integrated into a single replay buffer alongside agent experiences, directly addressing the distribution gap by providing contextually relevant expert guidance. The framework utilizes a modified Soft Actor-Critic (SAC) algorithm with a dual standard deviation policy network, outputting distinct σ_RL for exploration and σ_IL for imitation, allowing for adaptive balancing of these learning objectives. Furthermore, an uncertainty-based mechanism selectively invokes the online expert to guide the agent in challenging situations, making exploration more targeted and efficient. RLfOLD demonstrated superior performance in the CARLA NoCrash benchmark with significantly fewer resources, highlighting its effectiveness and efficiency in complex, real-time domains.

In conclusion, the intellectual trajectory of expert-guided exploration has evolved from static, offline demonstration datasets to dynamic, interactive, and online expert guidance. This progression effectively addresses critical challenges such as the distribution gap and sample inefficiency, particularly in safety-critical applications like autonomous driving and robotics. While methods leveraging offline demonstrations provide crucial initial boosts, the trend towards online and active expert interaction, exemplified by frameworks like RLfOLD, represents a significant step towards more robust, adaptive, and generalizable RL systems. Future research will likely focus on refining the mechanisms for generating and integrating online expert guidance, especially in real-world scenarios where "privileged

information" is unavailable, and developing more sophisticated expert models that can provide nuanced and context-aware interventions.

6.3 Exploration in Dynamic and Expanding Environments

The challenge of efficient exploration intensifies significantly in Reinforcement Learning (RL) when agents must operate in environments where the state and action spaces are not static but continually expand or evolve (13). Unlike traditional Markov Decision Processes (MDPs) that assume fixed state and action sets, real-world systems often undergo updates, introducing novel states, actions, or even entire sub-environments. This dynamic nature necessitates exploration strategies that can efficiently discover and integrate new information without incurring computationally prohibitive retraining costs or suffering from catastrophic forgetting of previously acquired knowledge. This specialized context forms a crucial subset of lifelong and continual learning, where agents must adapt to an unending stream of tasks or environmental changes (234; 90).

The concept of Incremental Reinforcement Learning (Incremental RL) has emerged to specifically address this challenge, focusing on how agents can efficiently adapt their policies to newly introduced states and actions. While lifelong learning broadly concerns sequential task learning and knowledge transfer (90; 47), Incremental RL distinguishes itself by tackling the explicit *expansion* of the underlying MDP structure. A seminal contribution by (192) formally defines Incremental RL and proposes the Dual-Adaptive ϵ-greedy Exploration (DAE) algorithm. This approach confronts the inherent inefficiency of standard exploration methods and the strong inductive biases that can arise from extensive prior learning, which often hinder adaptation to expanding environments. DAE employs a Meta Policy (Ψ) to adaptively determine a state-dependent ϵ by assessing the exploration convergence of a state (via TD-Error rate), thereby deciding *when* to explore. Concurrently, an Explorer (Φ) guides the agent to prioritize "least-tried" actions by estimating their relative frequencies, addressing *what* to explore. Crucially, DAE also incorporates strategies for incrementally adapting deep Q-networks by reusing trained policies and intelligently initializing new neurons and Q-values. This architectural flexibility,

combined with the dual-adaptive exploration mechanism, significantly reduces training overhead and enables robust adaptation to expanding search spaces without retraining from scratch.

The need for adaptive exploration in dynamic settings is also highlighted by research into non-stationary environments, which share some conceptual overlaps with expanding environments, though they differ mechanistically. For instance, (92) introduces Reactive Exploration to cope with continual domain shifts in lifelong reinforcement learning. This work demonstrates that policy-gradient methods benefit from strategies that track and react to non-stationarities, such as changes in reward functions or environmental dynamics, within an otherwise fixed state-action space. While Reactive Exploration focuses on adapting to *changes* in existing elements, DAE specifically addresses the *addition* of new states and actions. However, both underscore the broader necessity for exploration strategies that can actively adapt to environmental changes, rather than relying on static or pre-defined exploration schedules. Similarly, the importance of exploration for generalization to new, unseen environments, as explored by (139), aligns with the goals of Incremental RL. Their Exploration via Distributional Ensemble (EDE) method encourages exploration of states with high epistemic uncertainty, which is crucial for acquiring knowledge that aids decision-making in novel situations. While EDE aims to generalize within a potentially vast but fixed environment, DAE's focus is on efficiently integrating entirely new components into the agent's operational space, a distinction critical for truly open-ended learning systems (234).

Other advanced exploration techniques, such as those leveraging intrinsic motivation (51) or information gain maximization (34), aim to improve exploration efficiency by incentivizing agents to visit novel states or reduce uncertainty about the environment dynamics. For example, Variational Information Maximizing Exploration (VIME) (51) uses Bayesian neural networks to maximize information gain about environment dynamics. While powerful in static high-dimensional environments, their direct applicability and scalability to *continually expanding* state and action spaces present unique challenges. Prediction-error-based methods, like those underlying many intrinsic motivation

approaches, may struggle when the underlying dynamics model requires continuous architectural restructuring rather than just parameter updates. The sudden introduction of entirely new states or actions can render existing prediction models inaccurate or incomplete, requiring significant re-learning or architectural modifications that are not inherently handled by these methods. Count-based or density-based novelty methods, while effective for discovering unvisited regions, would need robust mechanisms to distinguish truly *new* states/actions from merely *unvisited* ones within the previously known space, and to efficiently update their density estimations for an ever-growing space. DAE's explicit focus on identifying and prioritizing newly available actions and states, overcoming the inductive bias from extensive prior learning, offers a more targeted solution to these architectural and knowledge-transfer challenges.

The integration of such adaptive exploration strategies with broader lifelong learning frameworks is a critical direction. Lifelong RL methods, such as model-based Bayesian approaches that estimate a hierarchical posterior to distill common task structures (90), offer mechanisms for backward transfer and efficient learning across related tasks. However, these often assume a fixed set of potential tasks or a stable underlying model structure. The challenge of Incremental RL lies in the dynamic *growth* of this structure, requiring not just transfer but also efficient architectural expansion and robust exploration of truly novel elements. Learned optimization methods, which meta-learn update rules to handle non-stationarity, plasticity loss, and exploration (167), offer a promising avenue by building adaptability directly into the learning process, potentially complementing DAE's specific exploration mechanisms.

In conclusion, the progression towards "Exploration in Dynamic and Expanding Environments" marks a crucial intellectual shift in Reinforcement Learning, moving beyond the static MDP assumption towards more realistic, evolving systems. While foundational exploration methods provide general tools, the work on Incremental RL, particularly the Dual-Adaptive ϵ-greedy Exploration (192), offers a targeted solution for environments where state and action spaces continually grow. Future research in this area will likely focus on extending these adaptive exploration strategies to more complex, partially observ-

able, or even multi-agent expanding environments, further enhancing the lifelong learning capabilities of RL agents in truly dynamic real-world scenarios, and integrating them more deeply with meta-learning and continual learning paradigms to address catastrophic forgetting and efficient knowledge transfer in ever-growing systems.

6.4 Safety-Aware Exploration

The exploration phase in reinforcement learning (RL) is critical for discovering optimal policies, yet in real-world, safety-critical applications, unconstrained exploration can lead to catastrophic outcomes and raise significant ethical concerns regarding accountability, fairness in decision-making under risk, and the potential for unforeseen negative side-effects in human-inhabited environments. This subsection delves into methods designed to ensure safety during exploration, navigating the inherent tension between the need for aggressive exploration to achieve optimal performance and the imperative to maintain safe operation. The problem is often formally cast within the framework of Constrained Markov Decision Processes (CMDPs), where an agent aims to maximize cumulative reward while simultaneously satisfying constraints on cumulative costs, such as safety violations (**?**). This formalism provides a robust theoretical foundation for developing algorithms that provide safety guarantees during the learning process.

Early efforts to integrate safety into RL exploration focused on establishing explicit boundaries and constraints. A foundational approach involves "safety layers" or "shielding," which act as guardians, restricting the agent's actions or states to predefined safe regions, thereby preventing the agent from entering hazardous situations during learning (**?**). While early works laid the groundwork, modern deep RL has seen significant advancements, notably with methods like those proposed by (**?**), which formally integrate shielding into deep RL agents. These methods enforce explicit safety constraints, ensuring exploration is guided within a permissible operational envelope, effectively mitigating the risk of catastrophic failures. However, a key limitation of static safety layers is their potential to be overly conservative, which can severely restrict the agent's exploration capabilities and prevent the discovery of truly optimal, yet initially unknown, safe be-

haviors. This conservativeness often stems from the difficulty of accurately predefining safe regions in complex, high-dimensional environments, or from worst-case assumptions made to guarantee safety.

Addressing the limitations of static constraints and the inherent conservativeness, more recent research has explored dynamic and learned safety mechanisms, often decoupling the concerns of task performance and safety. (45) introduced SEditor, a two-policy approach that learns a "safety editor policy" to transform potentially unsafe actions proposed by a utility-maximizing policy into safe ones. SEditor represents a conceptual shift from static, predefined shields to a learned, dynamic safety filter, allowing for more nuanced and state-dependent safety interventions. This method moves beyond simplified safety models, enabling the safety editor to learn complex safety functions, effectively acting as a dynamic shield. While SEditor demonstrates significantly lower constraint violation rates and maintains high utility, its effectiveness relies on the ability to train an accurate safety editor policy, which can be challenging in highly dynamic or unpredictable environments.

Further advancing dynamic safety, (7) introduced Recovery RL, which first leverages offline data to learn about constraint-violating zones. It then employs a task policy for reward maximization and a dedicated recovery policy that activates to guide the agent back to safety when constraint violation is likely. This dual-policy structure allows for more aggressive exploration by the task policy, relying on the learned recovery mechanism to prevent unsafe outcomes. Unlike SEditor, which modifies actions *before* execution, Recovery RL focuses on *recovering* from potentially unsafe trajectories, offering a different trade-off between proactive prevention and reactive correction. Recovery RL demonstrates superior efficiency in balancing task success and constraint adherence in complex, contact-rich manipulation tasks and image-based navigation, even on physical robots, by allowing the task policy greater freedom. Similarly, (200) proposed a method for safe RL with dead-ends avoidance and recovery. This approach constructs a boundary to discriminate between safe and unsafe states, equivalent to distinguishing dead-end states, thereby ensuring maximum safe exploration with minimal limitation. Like Recovery RL, it utilizes a decoupled framework with a task policy and a pre-trained recovery policy, along with a

safety critic, to ensure safe actions during online training. This strategy aims to achieve better task performance with fewer safety violations by carefully delineating the extent of guaranteed safe exploration, offering a more precise definition of "safe" exploration boundaries.

Another powerful paradigm for guaranteeing safety, particularly in continuous control systems, draws from control theory: Lyapunov stability and Control Barrier Functions (CBFs). These methods provide formal guarantees that a system's state will remain within a predefined safe set. Control Barrier Functions (CBFs) are functions of the state that define a safe set and whose derivatives can be constrained to render this set forward-invariant, thus preventing the agent from leaving it. (24) proposed a Barrier Lyapunov Function-based safe RL (BLF-SRL) algorithm for autonomous vehicles. This approach integrates BLF items into an optimized backstepping control method, constraining state variables within a designed safety region during learning. By decomposing optimal control with BLF items, it achieves safe exploration while learning adaptive uncertain items, ensuring both safety and performance optimization in safety-critical domains. While control-theoretic methods like BLF-SRL offer strong, often deterministic, safety guarantees, they typically require an accurate model of the system dynamics and specific assumptions about the environment, a limitation not shared by model-free, data-driven approaches like Recovery RL, which in turn provide only probabilistic safety assurances. Addressing this model dependency, (74) presented a method using Disturbance-Observer-Based Control Barrier Functions (DOB-CBFs). This approach avoids explicit model learning by leveraging disturbance observers to accurately estimate the pointwise value of uncertainty, which is then incorporated into a robust CBF condition. This allows for less conservative safety filters, especially during early learning, by effectively handling unknown disturbances without requiring extensive model training.

The inherent uncertainty in exploration necessitates risk-sensitive approaches. Building on Bayesian exploration principles (as discussed in Section 3.2), some approaches use uncertainty to define credible intervals for constraint satisfaction, leading to more principled conservative exploration. (157) introduced Lag-U, an uncertainty-augmented

Lagrangian safe RL algorithm for autonomous driving. This method uses deep ensembles to estimate epistemic uncertainty, which is then used to encourage exploration and learn a risk-sensitive policy by adaptively modifying safety constraints. Furthermore, it incorporates an intervention assurance mechanism based on quantified uncertainty to select safer actions during deployment. This allows for a better trade-off between efficiency and risk avoidance, preventing overly conservative policies by making safety decisions based on the agent's confidence in its predictions. Complementing this, (86) proposed a distributional reachability certificate (DRC) to address model uncertainties and characterize robust persistently safe states. Their framework builds a shield policy based on DRC to minimize constraint violations, especially during training, by considering the distribution of potential long-term constraint violations, thereby enhancing safety robustness against model uncertainty. Beyond explicit uncertainty estimation, some methods directly manipulate the learning process to balance reward and safety. (141) addressed the conflict between reward and safety gradients, proposing a soft switching policy optimization method. By analyzing and manipulating these gradients, their framework aims to achieve a better balance between optimizing for rewards and adhering to safety constraints, offering a more direct way to mitigate the inherent conflict compared to simply adding penalty terms in CMDPs.

In scenarios where some safety signals are available in a controlled environment, "guided safe exploration" can be employed to facilitate safe transfer learning. (202) proposed a method where a "guide" agent learns to explore safely without external rewards in a controlled environment where safety signals are available. This guide then helps compose a safe behavior policy for a "student" agent in a target task where safety violations are prohibited. The student policy is regularized towards the guide while it is unreliable, gradually reducing the guide's influence as it learns. This approach enables safe transfer learning and faster problem-solving in the target task, highlighting the utility of leveraging prior safety knowledge to bootstrap safe exploration in novel, safety-critical settings.

In summary, safety-aware exploration has evolved significantly, moving from static,

predefined safety layers and formal constraint satisfaction (CMDPs) to dynamic, learned recovery mechanisms, control-theoretic guarantees (BLFs, CBFs), and risk-sensitive approaches leveraging uncertainty quantification. The field continues to grapple with the fundamental tension between encouraging sufficient exploration for optimal learning and guaranteeing safety in real-world, safety-critical applications. Future directions will likely involve integrating more sophisticated formal verification techniques, developing robust and scalable uncertainty quantification for proactive safety prediction, designing intrinsically safe exploration strategies that adhere to ethical guidelines from the outset, and further refining gradient manipulation techniques to optimally balance conflicting objectives. The goal remains to enable autonomous agents to learn effectively without compromising human or environmental safety, ensuring responsible deployment in increasingly complex and sensitive domains.

7 Conclusion and Future Directions

7.1 Summary of Key Advancements

The journey of exploration in reinforcement learning (RL) reflects a continuous and sophisticated evolution, driven by the persistent challenge of balancing the acquisition of new information with the exploitation of known optimal actions. This review has traced a narrative arc from foundational heuristics to theoretically grounded algorithms, and subsequently to scalable intrinsic motivation techniques, culminating in adaptive, learned exploration strategies tailored for specialized contexts. This progression underscores the field's relentless innovation in addressing the increasing complexity of environments and the inherent difficulties of the exploration-exploitation dilemma.

Early attempts to navigate the exploration-exploitation trade-off, as discussed in Section 2, centered on basic heuristics like ϵ-greedy and the implicit exploration benefits of model-based planning architectures such as Dyna-Q. Concurrently, explicit exploration bonuses, often count-based, provided direct incentives for visiting novel states. While effective in tabular settings, these methods faced significant scalability challenges in high-

dimensional or continuous state spaces. For instance, early attempts to generalize count-based exploration to deep RL, such as mapping states to hash codes (2), provided a surprisingly strong baseline but highlighted the need for more robust, scalable novelty detection mechanisms.

A pivotal shift occurred with the development of theoretically grounded exploration strategies (Section 3), which sought to provide provable guarantees on learning efficiency. Principles like "optimism in the face of uncertainty" (OFU), embodied in algorithms like R-Max, offered PAC-MDP guarantees by optimistically valuing unexplored regions. Bayesian approaches, notably Thompson Sampling, provided a principled framework for managing uncertainty by maintaining posterior distributions over models or value functions. Further advancing this theoretical rigor, information-theoretic methods emerged, guiding exploration by maximizing knowledge gain about the environment or optimal policy. Techniques such as Variational Information Maximizing Exploration (VIME) (51) and MaxInfoRL (153) exemplify this, rewarding agents for transitions that significantly reduce uncertainty or improve the agent's internal model. This information-centric view, as surveyed by (34), moved beyond simple novelty to a more sophisticated understanding of learning progress. However, the computational complexity of maintaining accurate uncertainty estimates and the limitations of optimism in scenarios with partially observable rewards (227) often restricted their applicability to simpler environments.

The advent of deep reinforcement learning necessitated a paradigm shift, leading to the widespread adoption of intrinsic motivation techniques (Section 4). These methods generate internal reward signals to drive exploration, proving crucial in environments with sparse external rewards. Initial concepts of curiosity and novelty-seeking evolved into scalable approaches. Count-based methods were adapted for high-dimensional spaces using pseudo-counts and density models. A significant breakthrough came with prediction-error based curiosity, where agents are rewarded for encountering surprising or unpredictable observations, as seen in the Intrinsic Curiosity Module (ICM) (121; 29). This directed exploration towards aspects of the environment that improve the agent's internal dynamics model. To address the "noisy TV" problem, where agents are perpetually attracted

to uninformative stochasticity, robust intrinsic reward mechanisms like Random Network Distillation (RND) were developed, effectively filtering out irrelevant noise and focusing exploration on truly learnable novelty. The introduction of Random Latent Exploration (RLE) (187) further simplified this, offering deep exploration benefits by pursuing randomly sampled goals in a latent space without complex bonus calculations.

The field has continued to push towards more advanced and adaptive exploration strategies (Section 5). Hierarchical Reinforcement Learning (HRL) enabled structured exploration by decomposing tasks into sub-problems, allowing agents to explore at different levels of temporal abstraction. A particularly impactful direction is meta-learning for exploration, where agents learn *how* to explore effectively across diverse tasks. Algorithms like MAESN (6) demonstrate this by learning structured exploration strategies and latent exploration spaces from prior experience, injecting informed stochasticity into policies and outperforming task-agnostic methods. Furthermore, the development of decoupled exploration and exploitation policies (DEEP) (98) highlighted that separating these concerns can significantly boost sample efficiency, particularly in sparse reward settings. Integrated frameworks, such as Agent57, combine multiple techniques like RND, episodic memory, and adaptive meta-controllers to achieve state-of-the-art performance across a wide range of challenging environments. Diversity-driven exploration strategies (20) also contribute to preventing policies from being trapped in local optima by encouraging varied behaviors. Population-based and evolutionary methods offer a meta-level solution, leveraging multiple agents or meta-optimization to achieve more robust and global exploration in complex reward landscapes.

Finally, the application of these sophisticated exploration methods has expanded into specialized contexts (Section 6), demonstrating their versatility and practical impact. In offline reinforcement learning, where active exploration is impossible, the focus shifted to conservative exploration within fixed datasets, employing uncertainty estimation to avoid out-of-distribution actions (8; 205). Expert demonstrations and human feedback have been leveraged to guide exploration, significantly improving sample efficiency and overcoming sparse reward challenges in domains like robotics (1; 3). Safety-aware ex-

ploration has become critical for real-world applications, incorporating constraints and recovery policies to prevent hazardous actions (7). The challenges of dynamic and open-ended environments, which demand continuous adaptation and robust discovery, are also being addressed (234; 130). Emerging trends, such as the use of Decision-Pretrained Transformers (DPT) for in-context learning and adaptive exploration (128), hint at a future where powerful foundation models might inherently possess sophisticated exploration capabilities.

In summary, the field has progressed from simple, often undirected, exploration heuristics to theoretically grounded methods, then to scalable intrinsic motivation for deep RL, and finally to highly adaptive, learned, and integrated strategies. This continuous innovation has enabled RL agents to achieve unprecedented performance and robustness in increasingly complex and diverse tasks, while also addressing critical concerns like safety and data efficiency. The trajectory reflects a deep commitment to overcoming the persistent challenges of the exploration-exploitation dilemma, paving the way for more intelligent and autonomous learning systems.

7.2 Open Challenges and Theoretical Gaps

Despite significant advancements in reinforcement learning (RL) exploration, several fundamental challenges persist, particularly concerning scalability, robustness, sample efficiency, and the enduring gap between theoretical guarantees and practical applicability in complex, real-world settings. Addressing these issues is crucial for enabling RL agents to operate effectively in high-dimensional, stochastic, and open-ended environments.

A primary challenge lies in scaling exploration strategies to extremely high-dimensional or continuous state-action spaces, where traditional methods struggle due to the curse of dimensionality. Early count-based exploration, while effective in tabular settings (**?**), quickly becomes infeasible. Efforts to bridge this gap include methods that generalize counts to high-dimensional spaces, such as using hash codes (**?**) or pseudo-counts derived from density models (**?**). Concurrently, intrinsic motivation, often based on prediction error, emerged as a scalable heuristic. For instance, (**?**) utilized deep predictive models

to generate exploration bonuses in Atari games, demonstrating the potential of learned models to guide exploration in complex visual domains. However, these prediction-error methods, like the Intrinsic Curiosity Module (ICM) (**?**), often suffer from the "noisy TV problem," where uninformative stochasticity in the environment can generate spurious intrinsic rewards, leading to inefficient exploration. (**?**) addressed this by proposing Random Network Distillation (RND), which uses the prediction error of a fixed random network, proving more robust to environmental stochasticity and focusing exploration on learnable aspects of the environment. More recently, (**?**) proposed Implicit Posteriori Parameter Distribution Optimization (IPPDO) to improve exploration by modeling parameter uncertainty with implicit distributions, aiming for more flexible and efficient exploration in deep RL. Similarly, (**?**) introduced Robust Policy Optimization (RPO) to maintain high policy entropy throughout training, preventing premature convergence and ensuring sustained exploration in continuous action spaces.

Another persistent gap exists between methods offering strong theoretical guarantees and those providing practical scalability. Algorithms like R-Max (**?**) provide provable bounds on sample complexity and regret, but their reliance on explicit model learning or tabular representations limits their application to simpler, finite Markov Decision Processes (MDPs). Bridging this gap requires developing principled yet adaptable solutions for real-world complexity. (**?**) attempts this by proposing a computationally and statistically efficient model-based RL algorithm for specific model classes (Kernelized Nonlinear Regulators and linear MDPs) with polynomial sample complexity guarantees. In practical control applications, where accurate models are hard to obtain, methods like ModelPPO (**?**) integrate neural network models into actor-critic architectures for AUV path following, demonstrating improved performance over traditional and model-free RL by learning state transition functions. For dynamic systems like microgrids, (**?**) employs online RL with SARSA to adapt to uncertainties without relying on traditional mathematical models, prioritizing computational efficiency and real-time adaptability. Similarly, (**?**) proposes a lightweight, adaptive SAC algorithm for UAV path planning, which adjusts exploration probability dynamically to balance efficiency and generalization in resource-

constrained environments. These works highlight the ongoing tension between theoretical optimality and the need for practical, robust solutions in complex engineering domains.

The challenge of designing exploration strategies that are both sample-efficient and theoretically optimal across diverse tasks, especially in open-ended learning scenarios, remains largely unresolved. Meta-reinforcement learning (meta-RL) offers a promising avenue by learning exploration strategies from prior experience. (?) introduced MAESN (Model Agnostic Exploration with Structured Noise) to learn structured exploration strategies, which are more effective than task-agnostic noise. (?) further developed MetaCURE, an empowerment-driven exploration method for meta-RL that maximizes information gain for task identification in sparse-reward settings. Leveraging offline data can also significantly boost sample efficiency. (?) demonstrated that incorporating demonstrations can overcome exploration difficulties in sparse-reward robotics tasks, while (?) showed how to effectively integrate offline data into online RL with minimal modifications. Offline RL itself faces challenges with out-of-distribution actions, which (?) addresses with Uncertainty Weighted Actor-Critic (UWAC) by down-weighting contributions from uncertain state-action pairs. More recent work explores in-context learning with large transformer models: (?) introduced Decision-Pretrained Transformer (DPT), which can learn in-context exploration and conservatism from diverse datasets, generalizing to new tasks. Building on this, (?) proposed In-context Exploration-Exploitation (ICEE) to optimize this trade-off at inference time, enhancing efficiency. The development of benchmarks like Craftax (?) further underscores the current limitations of existing methods in achieving deep exploration, long-term planning, and continual adaptation required for truly open-ended learning. Cooperative exploration in multi-agent systems (? ?) and autonomous navigation (? ? ?) also represent significant real-world complexities demanding robust and adaptable exploration.

Finally, integrating safety constraints into exploration is a critical, yet often conflicting, requirement for real-world deployment. Extensive exploration, while necessary for learning, can lead to dangerous situations. (?) proposed Recovery RL, which decouples task and recovery policies and learns constraint-violating zones from offline data to safely

navigate this trade-off. (?) introduced SEditor, a two-policy approach where a safety editor policy transforms potentially unsafe actions into safe ones, achieving extremely low constraint violation rates. (?) advanced safe RL by identifying and avoiding dead-end states, providing a minimal limitation on safe exploration. In reward-free settings, (?) proposed a "guide" agent to learn safe exploration which then regularizes a "student" policy. For deployable safe RL, (?) developed Meta SAC-Lag, using meta-gradient optimization to automatically tune safety-related hyperparameters. Addressing model uncertainties for robust safety, (?) introduced a distributional reachability certificate for safe model-based RL. Furthermore, (?) applied distributionally robust RL for active signal pattern localization, enabling safe exploration in unfamiliar environments with limited data. These efforts highlight the ongoing challenge of balancing the need for exploration with stringent safety requirements, often requiring complex architectural designs or meta-learning approaches.

In conclusion, while the field has made substantial progress from tabular, theoretically-grounded methods to scalable, deep learning-based intrinsic motivation, significant open challenges remain. The persistent gap between methods with strong theoretical guarantees (often for simpler settings) and those providing practical scalability (often heuristic-driven) underscores the critical need for principled yet adaptable solutions. Future research must focus on developing exploration strategies that are robust to environmental stochasticity, sample-efficient across diverse tasks, capable of deep exploration in open-ended environments, and inherently safe for real-world deployment, potentially through hybrid approaches that combine the strengths of model-based reasoning, intrinsic motivation, and meta-learning with strong theoretical foundations.

7.3 Emerging Trends and Ethical Considerations

The frontier of reinforcement learning (RL) exploration is characterized by a dual pursuit: developing increasingly sophisticated agents capable of understanding and navigating complex, open-ended environments, and simultaneously ensuring these autonomous systems operate ethically and safely, particularly in human-interactive or safety-critical

domains. This subsection explores cutting-edge research directions, including the transformative integration of large foundation models, the development of truly general-purpose and adaptive exploration agents, the increasing focus on learning better representations, and the critical ethical implications of deploying such intelligent systems.

A pivotal emerging trend is the integration of **large foundation models (LFMs)**, such as Large Language Models (LLMs) and Vision-Language Models (VLMs), to imbue RL agents with more sophisticated world understanding, high-level planning capabilities, and common-sense priors. Traditional RL often struggles with extensive exploration in complex, semantically rich environments due to its limited grasp of underlying decision dynamics. LLMs, with their vast domain-specific knowledge, can serve as powerful prior action distributions, significantly reducing exploration and optimization complexity when integrated into RL frameworks through Bayesian inference methods (143). This approach can decrease the number of required samples by over 90

Complementing the rise of LFMs, there is an increasing focus on **learning better representations** to facilitate more informed and efficient exploration. Robust representations are crucial for defining novelty, quantifying uncertainty, and building accurate world models in high-dimensional observation spaces. While earlier methods like the simplified Intrinsic Curiosity Module (S-ICM) (121) and its predecessor ICM (**?**) leveraged prediction error in learned feature spaces to incentivize novelty, contemporary research pushes for more sophisticated self-supervised techniques that disentangle factors of variation and capture epistemic uncertainty. For example, the Actor-Model-Critic (AMC) architecture for Autonomous Underwater Vehicle (AUV) path-following explicitly learns the state transition function via a neural network, enabling the agent to anticipate environmental dynamics and guide exploration towards informative regions (129). Beyond explicit model learning, methods like Exploration via Distributional Ensemble (EDE) emphasize the importance of exploration for generalization, not just optimal policy finding (139). EDE encourages exploration of states with high epistemic uncertainty using an ensemble of Q-value distributions, implicitly relying on robust representations to quantify

this uncertainty effectively. Similarly, Decoupled Exploration and Exploitation Policies (DEEP) demonstrate that separating the task policy from the exploration policy can yield significant sample efficiency improvements in sparse environments, a benefit often amplified by representations that allow for meaningful novelty detection and uncertainty estimation (98). These approaches underscore that the quality of learned representations directly impacts an agent's ability to discern truly novel or uncertain aspects of an environment, leading to more directed and less wasteful exploration.

The drive towards **truly general-purpose exploration agents** capable of tackling open-ended problems is leading to more adaptive, robust, and scalable strategies. Rather than relying on fixed heuristics, recent work focuses on agents that can dynamically adjust their exploration behavior. Adaptive exploration strategies, such as those using multi-attribute decision-making based on information entropy and task decomposition, allow for more flexible and context-aware exploration (118). Further advancing this, ensemble learning schemes with explicit "exploration-to-exploitation (E2E) ratio control" via multiple Q-learning agents and adaptive decay functions enable more nuanced balancing of exploration and exploitation, crucial for real-world applications requiring continuous adaptation (159). The scalability and theoretical guarantees of exploration are also paramount for such agents. Thompson sampling-based methods, employing Langevin Monte Carlo (LMC) and approximate sampling, offer provably efficient and scalable exploration in deep RL with theoretical regret bounds, ensuring reliability in autonomous systems (148; 189). This extends to collaborative settings, where randomized exploration in cooperative multi-agent RL (MARL) with methods like CoopTS-PHE and CoopTS-LMC provides theoretical guarantees for regret bounds and communication complexity, essential for complex multi-agent environments (154). Moreover, simple yet effective strategies like Random Latent Exploration (RLE), which pursues randomly sampled goals in a latent space, demonstrate that deep exploration can be achieved without complex bonus calculations, promoting broader applicability as a general plug-in for existing RL algorithms (187). The concept of meta-learning how to explore is also gaining traction, with approaches like Learned Optimization for Plasticity, Exploration and Non-stationarity

(OPEN) meta-learning update rules that incorporate stochasticity for exploration, showing strong generalization across diverse environments and agent architectures (167). These advancements, alongside broader discussions on open-ended RL emphasizing hierarchical learning, intrinsic motivation, and unsupervised skill acquisition (234), signify a shift towards agents that can autonomously learn and adapt their exploration strategies across a wide spectrum of tasks.

Alongside these advancements in exploration capabilities, the **ethical considerations** surrounding autonomous exploration are gaining increasing prominence, especially in safety-critical or human-interactive environments. The inherent trial-and-error nature of RL exploration can lead to "bad decisions" that violate critical safety constraints, as highlighted in reviews of safe RL for power system control (213). This necessitates responsible development and deployment, emphasizing alignment with human values and safety standards. A direct response to this challenge is the "human-in-the-loop deep reinforcement learning (HL-DRL)" approach for optimal Volt/Var control in unbalanced distribution networks (212). This method allows human intervention to modify dangerous actions during offline training and integrates human guidance into the actor network's loss function, ensuring the learned policy adheres to operational constraints and human safety guidelines. Broader advancements in RL for autonomous systems also explicitly identify safety, dependability, and explainability as critical constraints that limit wide adoption (193). The imperative is to develop exploration strategies that not only discover optimal behaviors but do so within predefined safe regions, learn to recover from unsafe situations, and provide transparent decision-making processes. This ensures that the learning process does not lead to catastrophic outcomes and adheres to ethical considerations, bridging the gap between autonomous learning and responsible societal impact.

In conclusion, the field is rapidly advancing towards more intelligent, adaptive, and generalizable exploration strategies. This progress is driven by the transformative potential of large foundation models for high-level understanding and goal generation, the continuous refinement of learned representations for informed low-level novelty detection and uncertainty quantification, and the development of meta-learning approaches for truly

general-purpose agents. Simultaneously, the increasing power and autonomy of these systems amplify the imperative to address ethical implications, particularly in safety-critical domains. Future research must continue to bridge the gap between theoretical guarantees and practical deployment in highly dynamic real-world scenarios, further integrating human oversight, value alignment, and explainability into the design of autonomous exploration systems to ensure their beneficial and responsible societal impact.

References

References

[1] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, et al. (2017). *Overcoming Exploration in Reinforcement Learning with Demonstrations.* IEEE International Conference on Robotics and Automation.

[2] Haoran Tang, Rein Houthooft, Davis Foote, et al. (2016). *Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning.* Neural Information Processing Systems.

[3] Kimin Lee, Laura M. Smith, and P. Abbeel (2021). *PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training.* International Conference on Machine Learning.

[4] Junyan Hu, Hanlin Niu, J. Carrasco, et al. (2020). *Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning.* IEEE Transactions on Vehicular Technology.

[5] Bradly C. Stadie, S. Levine, and P. Abbeel (2015). *Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models.* arXiv.org.

[6] Abhishek Gupta, Russell Mendonca, Yuxuan Liu, et al. (2018). *Meta-Reinforcement Learning of Structured Exploration Strategies.* Neural Information Processing Systems.

[7] Brijen Thananjeyan, A. Balakrishna, Suraj Nair, et al. (2020). *Recovery RL: Safe Reinforcement Learning With Learned Recovery Zones.* IEEE Robotics and Automation Letters.

[8] Yue Wu, Shuangfei Zhai, Nitish Srivastava, et al. (2021). *Uncertainty Weighted Actor-Critic for Offline Reinforcement Learning.* International Conference on Machine Learning.

[9] Edoardo Conti, Vashisht Madhavan, F. Such, et al. (2017). *Improving Exploration in Evolution Strategies for Deep Reinforcement Learning via a Population of Novelty-Seeking Agents*. Neural Information Processing Systems.

[10] Younggyo Seo, Kimin Lee, Stephen James, et al. (2022). *Reinforcement Learning with Action-Free Pre-Training from Videos*. International Conference on Machine Learning.

[11] Ikechukwu Uchendu, Ted Xiao, Yao Lu, et al. (2022). *Jump-Start Reinforcement Learning*. International Conference on Machine Learning.

[12] Haoran Li, Qichao Zhang, and Dongbin Zhao (2020). *Deep Reinforcement Learning-Based Automatic Exploration for Navigation in Unknown Environment*. IEEE Transactions on Neural Networks and Learning Systems.

[13] Tianpei Yang, Hongyao Tang, Chenjia Bai, et al. (2021). *Exploration in Deep Reinforcement Learning: From Single-Agent to Multiagent Domain*. IEEE Transactions on Neural Networks and Learning Systems.

[14] Kimin Lee, Kibok Lee, Jinwoo Shin, et al. (2019). *Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning*. International Conference on Learning Representations.

[15] Junyu Zhang, Alec Koppel, A. S. Bedi, et al. (2020). *Variational Policy Gradient Method for Reinforcement Learning with General Utilities*. Neural Information Processing Systems.

[16] Jingru Chang, Dong Yu, Y. Hu, et al. (2022). *Deep Reinforcement Learning for Dynamic Flexible Job Shop Scheduling with Random Job Arrival*. Processes.

[17] Tianpei Yang, Hongyao Tang, Chenjia Bai, et al. (2021). *Exploration in Deep Reinforcement Learning: A Comprehensive Survey*. arXiv.org.

[18] Jinning Li, Chen Tang, M. Tomizuka, et al. (2022). *Hierarchical Planning Through*

Goal-Conditioned Offline Reinforcement Learning. IEEE Robotics and Automation Letters.

[19] Xi-Xi Liang, Katherine Shu, Kimin Lee, et al. (2022). *Reward Uncertainty for Exploration in Preference-based Reinforcement Learning*. International Conference on Learning Representations.

[20] Zhang-Wei Hong, Tzu-Yun Shann, Shih-Yang Su, et al. (2018). *Diversity-Driven Exploration Strategy for Deep Reinforcement Learning*. Neural Information Processing Systems.

[21] Nicklas Hansen, Yixin Lin, H. Su, et al. (2022). *MoDem: Accelerating Visual Model-Based Reinforcement Learning with Demonstrations*. International Conference on Learning Representations.

[22] Vitchyr H. Pong, Ashvin Nair, Laura M. Smith, et al. (2021). *Offline Meta-Reinforcement Learning with Online Self-Supervision*. International Conference on Machine Learning.

[23] Danyang Jia, Hao Guo, Z. Song, et al. (2021). *Local and global stimuli in reinforcement learning*. New Journal of Physics.

[24] Yuxiang Zhang, Xiaoling Liang, Dongyu Li, et al. (2022). *Barrier Lyapunov Function-Based Safe Reinforcement Learning for Autonomous Vehicles With Optimized Backstepping*. IEEE Transactions on Neural Networks and Learning Systems.

[25] Ron Dorfman, Idan Shenfeld, and Aviv Tamar (2021). *Offline Meta Reinforcement Learning - Identifiability Challenges and Effective Data Collection Strategies*. Neural Information Processing Systems.

[26] L. Tai, and Ming Liu (2016). *A robot exploration strategy based on Q-learning network*. International Conference on Real-time Computing and Robotics.

[27] Jarryd Martin, S. N. Sasikumar, Tom Everitt, et al. (2017). *Count-Based Exploration in Feature Space for Reinforcement Learning.* International Joint Conference on Artificial Intelligence.

[28] Julius Rückin, Liren Jin, and Marija Popovic (2021). *Adaptive Informative Path Planning Using Deep Reinforcement Learning for UAV-based Active Sensing.* IEEE International Conference on Robotics and Automation.

[29] Oleksii Zhelo, Jingwei Zhang, L. Tai, et al. (2018). *Curiosity-driven Exploration for Mapless Navigation with Deep Reinforcement Learning.* arXiv.org.

[30] Qilei Zhang, Jinying Lin, Q. Sha, et al. (2020). *Deep Interactive Reinforcement Learning for Path Following of Autonomous Underwater Vehicle.* IEEE Access.

[31] B. Mavrin, Shangtong Zhang, Hengshuai Yao, et al. (2019). *Distributional Reinforcement Learning for Efficient Exploration.* International Conference on Machine Learning.

[32] Gen Li, Laixi Shi, Yuxin Chen, et al. (2021). *Breaking the Sample Complexity Barrier to Regret-Optimal Model-Free Reinforcement Learning.* Neural Information Processing Systems.

[33] Pierre Schumacher, D. Haeufle, Dieter Büchler, et al. (2022). *DEP-RL: Embodied Exploration for Reinforcement Learning in Overactuated and Musculoskeletal Systems.* International Conference on Learning Representations.

[34] A. Aubret, L. Matignon, and S. Hassas (2022). *An Information-Theoretic Perspective on Intrinsic Motivation in Reinforcement Learning: A Survey.* Entropy.

[35] Xiaolei Yuan, Yiqun Pan, Jianrong Yang, et al. (2020). *Study on the application of reinforcement learning in the operation optimization of HVAC system.* Building Simulation.

[36] Shideh Rezaeifar, Robert Dadashi, Nino Vieillard, et al. (2021). *Offline Reinforcement Learning as Anti-Exploration.* AAAI Conference on Artificial Intelligence.

[37] Fatma M. Talaat (2022). *Effective deep Q-networks (EDQN) strategy for resource allocation based on optimized reinforcement learning algorithm.* Multimedia tools and applications.

[38] Xin Xin, Tiago Pimentel, Alexandros Karatzoglou, et al. (2022). *Rethinking Reinforcement Learning for Recommendation: A Prompt Perspective.* Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[39] Fengying Dang, Dong Chen, J. Chen, et al. (2022). *Event-Triggered Model Predictive Control With Deep Reinforcement Learning for Autonomous Driving.* IEEE Transactions on Intelligent Vehicles.

[40] A. Raffin, Jens Kober, and F. Stulp (2020). *Smooth Exploration for Robotic Reinforcement Learning.* Conference on Robot Learning.

[41] Xuhan Liu, K. Ye, H. V. van Vlijmen, et al. (2018). *An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor.* Journal of Cheminformatics.

[42] Qiyu Sun, Jinbao Fang, Weixing Zheng, et al. (2022). *Aggressive Quadrotor Flight Using Curiosity-Driven Reinforcement Learning.* IEEE transactions on industrial electronics (1982. Print).

[43] Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, et al. (2018). *Information-Directed Exploration for Deep Reinforcement Learning.* International Conference on Learning Representations.

[44] Dan Qiao, Ming Yin, Ming Min, et al. (2022). *Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost.* International Conference on Machine Learning.

[45] Haonan Yu, Wei Xu, and Haichao Zhang (2022). *Towards Safe Reinforcement Learning with a Safety Editor Policy.* Neural Information Processing Systems.

[46] Nathan Lambert, Markus Wulfmeier, William F. Whitney, et al. (2022). *The Challenges of Exploration for Offline Reinforcement Learning.* arXiv.org.

[47] Maciej Wołczyk, Michal Zajkac, Razvan Pascanu, et al. (2022). *Disentangling Transfer in Continual Reinforcement Learning.* Neural Information Processing Systems.

[48] Qingyu Qu, Kexin Liu, Wei Wang, et al. (2022). *Spacecraft Proximity Maneuvering and Rendezvous With Collision Avoidance Based on Reinforcement Learning.* IEEE Transactions on Aerospace and Electronic Systems.

[49] H. Sun, and Ling Ma (2020). *Generative Design by Using Exploration Approaches of Reinforcement Learning in Density-Based Structural Topology Optimization.* Designs.

[50] Yuechuan Tao, Jing Qiu, Shuying Lai, et al. (2022). *A Human-Machine Reinforcement Learning Method for Cooperative Energy Management.* IEEE Transactions on Industrial Informatics.

[51] Rein Houthooft, Xi Chen, Yan Duan, et al. (2016). *Curiosity-driven Exploration in Deep Reinforcement Learning via Bayesian Neural Networks.* arXiv.org.

[52] Tianyu Shi, Dong Chen, Kaian Chen, et al. (2021). *Offline Reinforcement Learning for Autonomous Driving with Safety and Exploration Enhancement.* arXiv.org.

[53] Zhiwei Li, Yu Lu, Xi Li, et al. (2021). *UAV Networks Against Multiple Maneuvering Smart Jamming With Knowledge-Based Reinforcement Learning.* IEEE Internet of Things Journal.

[54] Jiabin Liu, Chengliang Chai, Yuyu Luo, et al. (2022). *Feature Augmentation with Reinforcement Learning.* IEEE International Conference on Data Engineering.

[55] Hangkai Hu, Shiji Song, and C. L. Phillip Chen (2019). *Plume Tracing via Model-Free Reinforcement Learning Method.* IEEE Transactions on Neural Networks and Learning Systems.

[56] Yutong Wang, Ke Xue, and Chaojun Qian (2022). *Evolutionary Diversity Optimization with Clustering-based Selection for Reinforcement Learning.* International Conference on Learning Representations.

[57] Chao Yu, Xinyi Yang, Jiaxuan Gao, et al. (2021). *Learning Efficient Multi-Agent Cooperative Visual Exploration.* European Conference on Computer Vision.

[58] Changdong Zheng, Fangfang Xie, Tingwei Ji, et al. (2022). *Data-efficient deep reinforcement learning with expert demonstration for active flow control.* The Physics of Fluids.

[59] Zhengyu Yang, Kan Ren, Xufang Luo, et al. (2022). *Towards Applicable Reinforcement Learning: Improving the Generalization and Sample Efficiency with Policy Ensemble.* International Joint Conference on Artificial Intelligence.

[60] Yijun Yang, J. Jiang, Tianyi Zhou, et al. (2022). *Pareto Policy Pool for Model-based Offline Reinforcement Learning.* International Conference on Learning Representations.

[61] Yuqi Liu, Po Gao, Change Zheng, et al. (2022). *A Deep Reinforcement Learning Strategy Combining Expert Experience Guidance for a Fruit-Picking Manipulator.* Electronics.

[62] Zhongni Hou, Xiaolong Jin, Zixuan Li, et al. (2021). *Rule-Aware Reinforcement Learning for Knowledge Graph Reasoning.* Findings.

[63] Sahin Lale, K. Azizzadenesheli, B. Hassibi, et al. (2020). *Reinforcement Learning with Fast Stabilization in Linear Dynamical Systems.* International Conference on Artificial Intelligence and Statistics.

[64] Fabian Otto, Onur Çelik, Hongyi Zhou, et al. (2022). *Deep Black-Box Reinforcement Learning with Movement Primitives.* Conference on Robot Learning.

[65] Ayub I. Lakhani, Myisha A. Chowdhury, and Qiugang Lu (2021). *Stability-Preserving Automatic Tuning of PID Control with Reinforcement Learning.* Complex Engineering Systems.

[66] Yixin Huang, Shufan Wu, Z. Mu, et al. (2020). *A Multi-agent Reinforcement Learning Method for Swarm Robots in Space Collaborative Exploration.* 2020 6th International Conference on Control, Automation and Robotics (ICCAR).

[67] Wang Yuan, Z. Xiwen, Zhou Rong, et al. (2022). *Research on UCAV Maneuvering Decision Method Based on Heuristic Reinforcement Learning.* Computational Intelligence and Neuroscience.

[68] Abu Jafar Md. Muzahid, Syafiq Fauzi Bin Kamarulzaman, Md. Arafatur Rahman, et al. (2022). *Deep Reinforcement Learning-Based Driving Strategy for Avoidance of Chain Collisions and Its Safety Efficiency Analysis in Autonomous Vehicles.* IEEE Access.

[69] Chi Zhang, S. Kuppannagari, and V. Prasanna (2022). *Safe Building HVAC Control via Batch Reinforcement Learning.* IEEE Transactions on Sustainable Computing.

[70] J. E. Sierra-García, and Matilde Santos (2020). *Exploring Reward Strategies for Wind Turbine Pitch Control by Reinforcement Learning.* Applied Sciences.

[71] Wei Han, Fang Guo, and Xi-chao Su (2019). *A Reinforcement Learning Method for a Hybrid Flow-Shop Scheduling Problem.* Algorithms.

[72] K. P. Wabersich, and M. Zeilinger (2018). *Safe exploration of nonlinear dynamical systems: A predictive safety filter for reinforcement learning.* arXiv.org.

[73] Hippolyte Bourel, Odalric-Ambrym Maillard, and M. S. Talebi (2020). *Tightening Exploration in Upper Confidence Reinforcement Learning.* International Conference on Machine Learning.

[74] Yikun Cheng, Pan Zhao, and N. Hovakimyan (2022). *Safe and Efficient Reinforcement Learning using Disturbance-Observer-Based Control Barrier Functions.* Conference on Learning for Dynamics Control.

[75] Glenn Ceusters, L. R. Camargo, R. Franke, et al. (2022). *Safe reinforcement learning for multi-energy management systems with known constraint functions.* Energy and AI.

[76] C. Stanton, and J. Clune (2018). *Deep Curiosity Search: Intra-Life Exploration Improves Performance on Challenging Deep Reinforcement Learning Problems.* arXiv.org.

[77] Geoffrey Cideron, Thomas Pierrot, Nicolas Perrin, et al. (2020). *QD-RL: Efficient Mixing of Quality and Diversity in Reinforcement Learning.* arXiv.org.

[78] Xiangyu Liu, and Ying Tan (2022). *Feudal Latent Space Exploration for Coordinated Multi-Agent Reinforcement Learning.* IEEE Transactions on Neural Networks and Learning Systems.

[79] Daesol Cho, Jigang Kim, and H. J. Kim (2022). *Unsupervised Reinforcement Learning for Transferable Manipulation Skill Discovery.* IEEE Robotics and Automation Letters.

[80] Jin Zhang, Jianhao Wang, Hao Hu, et al. (2020). *MetaCURE: Meta Reinforcement Learning with Empowerment-Driven Exploration.* International Conference on Machine Learning.

[81] Yuda Song, and Wen Sun (2021). *PC-MLP: Model-based Reinforcement Learning with Policy Cover Guided Exploration.* International Conference on Machine Learning.

[82] Xiucheng Wang, Longfei Ma, Hao Li, et al. (2022). *Digital Twin-Assisted Efficient Reinforcement Learning for Edge Task Scheduling.* IEEE Vehicular Technology Conference.

[83] Sen Lin, Jialin Wan, Tengyu Xu, et al. (2022). *Model-Based Offline Meta-Reinforcement Learning with Regularization.* International Conference on Learning Representations.

[84] Ziqian Zhang, Yulei Liu, Shengcheng Yu, et al. (2022). *UniRLTest: universal platform-independent testing with reinforcement learning via image understanding.* International Symposium on Software Testing and Analysis.

[85] Tong Zhou, Letian Wang, Ruobing Chen, et al. (2022). *Accelerating Reinforcement Learning for Autonomous Driving Using Task-Agnostic and Ego-Centric Motion Skills.* IEEE/RJS International Conference on Intelligent RObots and Systems.

[86] Dongjie Yu, Wenjun Zou, Yujie Yang, et al. (2022). *Safe Model-Based Reinforcement Learning With an Uncertainty-Aware Reachability Certificate.* IEEE Transactions on Automation Science and Engineering.

[87] Christopher Xie, S. Patil, T. Moldovan, et al. (2015). *Model-based reinforcement learning with parametrized physical models and optimism-driven exploration.* IEEE International Conference on Robotics and Automation.

[88] Yu Zhang, Peixiang Cai, Changyong Pan, et al. (2019). *Multi-Agent Deep Reinforcement Learning-Based Cooperative Spectrum Sensing With Upper Confidence Bound Exploration.* IEEE Access.

[89] Zhenning Wu, Yiming Deng, Jinhai Liu, et al. (2021). *A Reinforcement Learning-Based Reconstruction Method for Complex Defect Profiles in MFL Inspection.* IEEE Transactions on Instrumentation and Measurement.

[90] Haotian Fu, Shangqun Yu, Michael S. Littman, et al. (2022). *Model-based Lifelong Reinforcement Learning with Bayesian Exploration.* Neural Information Processing Systems.

[91] Arquímides Méndez-Molina, E. Morales, and L. Sucar (2022). *Causal Discovery and Reinforcement Learning: A Synergistic Integration.* European Workshop on Probabilistic Graphical Models.

[92] C. Steinparz, Thomas Schmied, Fabian Paischer, et al. (2022). *Reactive Exploration to Cope with Non-Stationarity in Lifelong Reinforcement Learning.* CoLLAs.

[93] Md Masudur Rahman, and Yexiang Xue (2022). *Robust Policy Optimization in Deep Reinforcement Learning.* arXiv.org.

[94] Jingyi Xu, Zirui Li, Li Gao, et al. (2022). *A Comparative Study of Deep Reinforcement Learning-based Transferable Energy Management Strategies for Hybrid Electric Vehicles.* 2022 IEEE Intelligent Vehicles Symposium (IV).

[95] Kyunghyun Lee, Byeong-uk Lee, Ukcheol Shin, et al. (2020). *An Efficient Asynchronous Method for Integrating Evolutionary and Gradient-based Policy Search.* Neural Information Processing Systems.

[96] Ziniu Li, Yingru Li, Yushun Zhang, et al. (2022). *HyperDQN: A Randomized Exploration Method for Deep Reinforcement Learning.* International Conference on Learning Representations.

[97] Zhihai Wang, Taoxing Pan, Qi Zhou, et al. (2022). *Efficient Exploration in Resource-Restricted Reinforcement Learning.* AAAI Conference on Artificial Intelligence.

[98] William F. Whitney, Michael Bloesch, Jost Tobias Springenberg, et al. (2021). *Decoupled Exploration and Exploitation Policies for Sample-Efficient Reinforcement Learning.* Unpublished manuscript.

[99] Karush Suri (2022). *Off-Policy Evolutionary Reinforcement Learning with Maximum Mutations.* Adaptive Agents and Multi-Agent Systems.

[100] Bo Xin, Haixu Yu, You Qin, et al. (2020). *Exploration Entropy for Reinforcement Learning.* Unpublished manuscript.

[101] Guillaume Matheron, Nicolas Perrin, and Olivier Sigaud (2020). *PBCS : Efficient Exploration and Exploitation Using a Synergy between Reinforcement Learning and Motion Planning.* International Conference on Artificial Neural Networks.

[102] Yiqin Yang, Haotian Hu, Wenzhe Li, et al. (2022). *Flow to Control: Offline Reinforcement Learning with Lossless Primitive Discovery.* AAAI Conference on Artificial Intelligence.

[103] Zheng Wu, Yichen Xie, Wenzhao Lian, et al. (2022). *Zero-Shot Policy Transfer with Disentangled Task Representation of Meta-Reinforcement Learning.* IEEE International Conference on Robotics and Automation.

[104] Samuel Kessler, Piotr Milo's, Jack Parker-Holder, et al. (2022). *The Surprising Effectiveness of Latent World Models for Continual Reinforcement Learning.* arXiv.org.

[105] A. Raffin, and F. Stulp (2020). *Generalized State-Dependent Exploration for Deep Reinforcement Learning in Robotics.* arXiv.org.

[106] Kai-En Yang, Chia-Yu Tsai, Hung-Hao Shen, et al. (2020). *Trust-Region Method with Deep Reinforcement Learning in Analog Design Space Exploration.* Design Automation Conference.

[107] Jiayi Liu, Gang Wang, Xiangke Guo, et al. (2022). *Deep Reinforcement Learning Task Assignment Based on Domain Knowledge.* IEEE Access.

[108] A. Kamalova, Suk-Gyu Lee, and Soon-H. Kwon (2022). *Occupancy Reward-Driven Exploration with Deep Reinforcement Learning for Mobile Robot System.* Applied Sciences.

[109] Haichao Zhang, Wei Xu, and Haonan Yu (2022). *Generative Planning for Temporally Coordinated Exploration in Reinforcement Learning.* International Conference on Learning Representations.

[110] Wenli Li, Yousong Zhang, Xiaohui Shi, et al. (2022). *A Decision-Making Strategy for Car Following Based on Naturalist Driving Data via Deep Reinforcement Learning.* Italian National Conference on Sensors.

[111] Lanxiao Huang, Tyler Cody, Christopher Redino, et al. (2022). *Exposing Surveillance Detection Routes via Reinforcement Learning, Attack Graphs, and Cyber Terrain*. International Conference on Machine Learning and Applications.

[112] Aditya Modi, and Ambuj Tewari (2019). *No-regret Exploration in Contextual Reinforcement Learning*. Conference on Uncertainty in Artificial Intelligence.

[113] Jiahe Shi, Yali Li, and Shengjin Wang (2021). *Partial Off-policy Learning: Balance Accuracy and Diversity for Human-Oriented Image Captioning*. IEEE International Conference on Computer Vision.

[114] Hengzhe Zhang, and Aimin Zhou (2021). *RL-GEP: Symbolic Regression via Gene Expression Programming and Reinforcement Learning*. IEEE International Joint Conference on Neural Network.

[115] Dujia Yang, Xiaowei Qin, Xiaodong Xu, et al. (2020). *Sample Efficient Reinforcement Learning Method via High Efficient Episodic Memory*. IEEE Access.

[116] Zhenshan Bing, Christian Lemke, Zhuangyi Jiang, et al. (2019). *Energy-Efficient Slithering Gait Exploration for a Snake-like Robot based on Reinforcement Learning*. International Joint Conference on Artificial Intelligence.

[117] Songan Zhang, H. Peng, S. Nageshrao, et al. (2019). *Discretionary Lane Change Decision Making using Reinforcement Learning with Model-Based Exploration*. International Conference on Machine Learning and Applications.

[118] Chunyang Hu, and Meng Xu (2020). *Adaptive Exploration Strategy With Multi-Attribute Decision-Making for Reinforcement Learning*. IEEE Access.

[119] K. N. Kumar, Irfan Essa, and Sehoon Ha (2021). *Graph-based Cluttered Scene Generation and Interactive Exploration using Deep Reinforcement Learning*. IEEE International Conference on Robotics and Automation.

[120] Erick Asiain, J. Clempner, and A. Poznyak (2018). *Controller exploitation-exploration reinforcement learning architecture for computing near-optimal policies.* Soft Computing - A Fusion of Foundations, Methodologies and Applications.

[121] Boyao Li, Tao Lu, Jiayi Li, et al. (2019). *Curiosity-Driven Exploration for Off-Policy Reinforcement Learning Methods*.* IEEE International Conference on Robotics and Biomimetics.

[122] Hao Sun, Ziping Xu, Yuhang Song, et al. (2020). *Zeroth-Order Supervised Policy Improvement.* arXiv.org.

[123] Yixuan Su, Deng Cai, Yan Wang, et al. (2020). *Stylistic Dialogue Generation via Information-Guided Reinforcement Learning Strategy.* arXiv.org.

[124] Hui Liu, Zhen Zhang, and Dongqing Wang (2020). *WRFMR: A Multi-Agent Reinforcement Learning Method for Cooperative Tasks.* IEEE Access.

[125] Philip J. Ball, Laura M. Smith, Ilya Kostrikov, et al. (2023). *Efficient Online Reinforcement Learning with Offline Data.* International Conference on Machine Learning.

[126] Qing-ran Meng, Sheharyar Hussain, Fengzhang Luo, et al. (2025). *An Online Reinforcement Learning-Based Energy Management Strategy for Microgrids With Centralized Control.* IEEE transactions on industry applications.

[127] Shihan Dou, Yan Liu, Haoxiang Jia, et al. (2024). *StepCoder: Improve Code Generation with Reinforcement Learning from Compiler Feedback.* arXiv.org.

[128] Jonathan Lee, Annie Xie, Aldo Pacchiano, et al. (2023). *Supervised Pretraining Can Learn In-Context Reinforcement Learning.* Neural Information Processing Systems.

[129] D. Ma, Xi Chen, Weihao Ma, et al. (2024). *Neural Network Model-Based Reinforcement Learning Control for AUV 3-D Path Following.* IEEE Transactions on Intelligent Vehicles.

[130] Michael Matthews, Michael Beukman, Benjamin Ellis, et al. (2024). *Craftax: A Lightning-Fast Benchmark for Open-Ended Reinforcement Learning*. International Conference on Machine Learning.

[131] Meng Xi, Huiao Dai, Jingyi He, et al. (2024). *A Lightweight Reinforcement-Learning-Based Real-Time Path-Planning Method for Unmanned Aerial Vehicles*. IEEE Internet of Things Journal.

[132] Jing Zhang, Jian-Lin Ren, Yani Cui, et al. (2024). *Multi-USV Task Planning Method Based on Improved Deep Reinforcement Learning*. IEEE Internet of Things Journal.

[133] Zhiheng Xi, Wenxiang Chen, Boyang Hong, et al. (2024). *Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning*. International Conference on Machine Learning.

[134] Nan Cheng, Xiucheng Wang, Zan Li, et al. (2024). *Toward Enhanced Reinforcement Learning-Based Resource Management via Digital Twin: Opportunities, Applications, and Challenges*. IEEE Network.

[135] Yihao Sun, Jiajin Zhang, Chengxing Jia, et al. (2023). *Model-Bellman Inconsistency for Model-based Offline Reinforcement Learning*. International Conference on Machine Learning.

[136] Guowei Xu, Ruijie Zheng, Yongyuan Liang, et al. (2023). *DrM: Mastering Visual Reinforcement Learning through Dormant Ratio Minimization*. International Conference on Learning Representations.

[137] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, et al. (2025). *R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning*. arXiv.org.

[138] Guanxing Lu, Wenkai Guo, Chubin Zhang, et al. (2025). *VLA-RL: Towards Masterful and General Robotic Manipulation with Scalable Reinforcement Learning*. arXiv.org.

[139] Yiding Jiang, J. Z. Kolter, and R. Raileanu (2023). *On the Importance of Exploration for Generalization in Reinforcement Learning.* Neural Information Processing Systems.

[140] Ruoqing Zhang, Ziwei Luo, Jens Sjölund, et al. (2024). *Entropy-regularized Diffusion Policy with Q-Ensembles for Offline Reinforcement Learning.* Neural Information Processing Systems.

[141] Shangding Gu, Bilgehan Sel, Yuhao Ding, et al. (2024). *Balance Reward and Safety Optimization for Safe Reinforcement Learning: A Perspective of Gradient Manipulation.* AAAI Conference on Artificial Intelligence.

[142] Runyu Ma, Jelle Luijkx, Zlatan Ajanović, et al. (2024). *ExploRLLM: Guiding Exploration in Reinforcement Learning with Large Language Models.* IEEE International Conference on Robotics and Automation.

[143] Xue Yan, Yan Song, Xidong Feng, et al. (2024). *Efficient Reinforcement Learning with Large Language Model Priors.* arXiv.org.

[144] Yinda Chen, Wei Huang, Shenglong Zhou, et al. (2023). *Self-Supervised Neuron Segmentation with Multi-Agent Reinforcement Learning.* International Joint Conference on Artificial Intelligence.

[145] Kuo Li, Xinze Jin, Qing-Shan Jia, et al. (2024). *An OCBA-Based Method for Efficient Sample Collection in Reinforcement Learning.* IEEE Transactions on Automation Science and Engineering.

[146] Tao Huang, Kai Chen, Bin Li, et al. (2023). *Demonstration-Guided Reinforcement Learning with Efficient Exploration for Task Automation of Surgical Robot.* IEEE International Conference on Robotics and Automation.

[147] Xinze Jin, Kuo Li, and Qing-Shan Jia (2024). *Constrained reinforcement learning with statewise projection: a control barrier function approach.* Science China Information Sciences.

[148] Haque Ishfaq, Qingfeng Lan, Pan Xu, et al. (2023). *Provable and Practical: Efficient Exploration in Reinforcement Learning via Langevin Monte Carlo.* International Conference on Learning Representations.

[149] Siyuan Guo, Lixin Zou, Hechang Chen, et al. (2024). *Sample Efficient Offline-to-Online Reinforcement Learning.* IEEE Transactions on Knowledge and Data Engineering.

[150] Anja Surina, Amin Mansouri, Lars Quaedvlieg, et al. (2025). *Algorithm Discovery With LLMs: Evolutionary Search Meets Reinforcement Learning.* arXiv.org.

[151] Onur Celik, Zechu Li, Denis Blessing, et al. (2025). *DIME:Diffusion-Based Maximum Entropy Reinforcement Learning.* arXiv.org.

[152] Hean Hua, and Yongchun Fang (2023). *A Novel Reinforcement Learning-Based Robust Control Strategy for a Quadrotor.* IEEE transactions on industrial electronics (1982. Print).

[153] Bhavya Sukhija, Stelian Coros, Andreas Krause, et al. (2024). *MaxInfoRL: Boosting exploration in reinforcement learning through information gain maximization.* International Conference on Learning Representations.

[154] Hao-Lun Hsu, Weixin Wang, Miroslav Pajic, et al. (2024). *Randomized Exploration in Cooperative Multi-Agent Reinforcement Learning.* Neural Information Processing Systems.

[155] Ziyi Wang, Xinran Li, Luoyang Sun, et al. (2024). *Learning State-Specific Action Masks for Reinforcement Learning.* Algorithms.

[156] S. Ghamari, Mojtaba Hajihosseini, D. Habibi, et al. (2024). *Design of an Adaptive Robust PI Controller for DC/DC Boost Converter Using Reinforcement-Learning Technique and Snake Optimization Algorithm.* IEEE Access.

[157] Zhengran Zhang, Qi Liu, Yanjie Li, et al. (2024). *Safe Reinforcement Learning in*

Autonomous Driving With Epistemic Uncertainty Estimation. IEEE transactions on intelligent transportation systems (Print).

[158] Zhenwen Dai, Federico Tomasi, and Sina Ghiassian (2024). *In-context Exploration-Exploitation for Reinforcement Learning*. International Conference on Learning Representations.

[159] Bin Shuai, Min Hua, Yanfei Li, et al. (2025). *Optimal Energy Management of Plug-In Hybrid Electric Vehicles Through Ensemble Reinforcement Learning With Exploration-to-Exploitation Ratio Control*. IEEE Transactions on Intelligent Vehicles.

[160] Roland Stolz, Hanna Krasowski, Jakob Thumm, et al. (2024). *Excluding the Irrelevant: Focusing Reinforcement Learning through Continuous Action Masking*. Neural Information Processing Systems.

[161] Martin Tappler, Andrea Pferscher, B. Aichernig, et al. (2024). *Learning and Repair of Deep Reinforcement Learning Policies from Fuzz-Testing Data*. International Conference on Software Engineering.

[162] Zohar Rimon, Tom Jurgenson, Orr Krupnik, et al. (2024). *MAMBA: an Effective World Model Approach for Meta-Reinforcement Learning*. International Conference on Learning Representations.

[163] Juan R. Terven (2025). *Deep Reinforcement Learning: A Chronological Overview and Methods*. Applied Informatics.

[164] Hao-Hsiang Hsiao, Yi-Chen Lu, Pruek Vanna-iampikul, et al. (2024). *FastTuner: Transferable Physical Design Parameter Optimization using Fast Reinforcement Learning*. ACM International Symposium on Physical Design.

[165] R. Kakooee, and B. Dillenburger (2024). *Reimagining space layout design through deep reinforcement learning*. Journal of Computational Design and Engineering.

[166] Rafael Rafailov, K. Hatch, Anikait Singh, et al. (2024). *D5RL: Diverse Datasets for Data-Driven Deep Reinforcement Learning*. RLJ.

[167] A. D. Goldie, Chris Lu, Matthew Jackson, et al. (2024). *Can Learned Optimization Make Reinforcement Learning Less Difficult?*. Neural Information Processing Systems.

[168] Daniel Coelho, Miguel Oliveira, and Vitor Santos (2024). *RLfOLD: Reinforcement Learning from Online Demonstrations in Urban Autonomous Driving*. AAAI Conference on Artificial Intelligence.

[169] Yanjun Huang, Yuxiao Gu, Kang Yuan, et al. (2024). *Human Knowledge Enhanced Reinforcement Learning for Mandatory Lane-Change of Autonomous Vehicles in Congested Traffic*. IEEE Transactions on Intelligent Vehicles.

[170] Daesol Cho, Seungjae Lee, and H. J. Kim (2023). *Outcome-directed Reinforcement Learning by Uncertainty Temporal Distance-Aware Curriculum Goal Generation*. International Conference on Learning Representations.

[171] Yucheng Shi, Wenhao Yu, Zaitang Li, et al. (2025). *MobileGUI-RL: Advancing Mobile GUI Agent through Reinforcement Learning in Online Environment*. arXiv.org.

[172] Ge Li, Hongyi Zhou, Dominik Roth, et al. (2024). *Open the Black Box: Step-based Policy Updates for Temporally-Correlated Episodic Reinforcement Learning*. International Conference on Learning Representations.

[173] Majid Ghasemi, Amir Hossein Moosavi, and Dariush Ebrahimi (2024). *A Comprehensive Survey of Reinforcement Learning: From Algorithms to Practical Challenges*. Unpublished manuscript.

[174] Zhen-yu Liu, Ke Wang, Daxin Liu, et al. (2023). *A Motion Planning Method for Visual Servoing Using Deep Reinforcement Learning in Autonomous Robotic Assembly*. IEEE/ASME transactions on mechatronics.

[175] Zhiwei Shang, Renxing Li, Chunhuang Zheng, et al. (2023). *Relative Entropy Regularized Sample-Efficient Reinforcement Learning With Continuous Actions.* IEEE Transactions on Neural Networks and Learning Systems.

[176] Zeyang Liu, Lipeng Wan, Xinrui Yang, et al. (2024). *Imagine, Initialize, and Explore: An Effective Exploration Method in Multi-Agent Reinforcement Learning.* AAAI Conference on Artificial Intelligence.

[177] Y. Kantaros, and Jun Wang (2024). *Sample-Efficient Reinforcement Learning With Temporal Logic Objectives: Leveraging the Task Specification to Guide Exploration.* IEEE Transactions on Automatic Control.

[178] Siqi Li, Jun Chen, Shanqi Liu, et al. (2024). *MCMC: Multi-Constrained Model Compression via One-Stage Envelope Reinforcement Learning.* IEEE Transactions on Neural Networks and Learning Systems.

[179] Dony Ang, Cyril Rakovski, and H. Atamian (2024). *De Novo Drug Design Using Transformer-Based Machine Translation and Reinforcement Learning of an Adaptive Monte Carlo Tree Search.* Pharmaceuticals.

[180] Jangsaeng Kim, Wonjun Shin, Jiyong Yim, et al. (2024). *Toward Optimized In-Memory Reinforcement Learning: Leveraging 1/f Noise of Synaptic Ferroelectric Field-Effect-Transistors for Efficient Exploration.* Advanced Intelligent Systems.

[181] Chengzhong Ma, Deyu Yang, Tianyu Wu, et al. (2024). *Improving Offline Reinforcement Learning With in-Sample Advantage Regularization for Robot Manipulation.* IEEE Transactions on Neural Networks and Learning Systems.

[182] Zhihong Ge, Xingshuo Li, Fei Xu, et al. (2024). *An Improved Distributed Maximum Power Point Tracking Technique in Photovoltaic Systems Based on Reinforcement Learning Algorithm.* IEEE Journal of Emerging and Selected Topics in Industrial Electronics.

[183] Feiyu Lu, Mengyu Chen, Hsiang Hsu, et al. (2024). *Adaptive 3D UI Placement in Mixed Reality Using Deep Reinforcement Learning.* CHI Extended Abstracts.

[184] Jianyong Yuan, Peiyu Wang, Junjie Ye, et al. (2023). *EasySO: Exploration-enhanced Reinforcement Learning for Logic Synthesis Sequence Optimization and a Comprehensive RL Environment.* 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD).

[185] Bowen Zheng, and Ran Cheng (2023). *Rethinking Population-assisted Off-policy Reinforcement Learning.* Annual Conference on Genetic and Evolutionary Computation.

[186] Tianfu Wang, Qilin Fan, Chao Wang, et al. (2024). *FlagVNE: A Flexible and Generalizable Reinforcement Learning Framework for Network Resource Allocation.* International Joint Conference on Artificial Intelligence.

[187] Srinath Mahankali, Zhang-Wei Hong, Ayush Sekhari, et al. (2024). *Random Latent Exploration for Deep Reinforcement Learning.* International Conference on Machine Learning.

[188] Youngsik Yoon, Gangbok Lee, Sungsoo Ahn, et al. (2024). *Breadth-First Exploration on Adaptive Grid for Reinforcement Learning.* International Conference on Machine Learning.

[189] Haque Ishfaq, Yixin Tan, Yu Yang, et al. (2024). *More Efficient Randomized Exploration for Reinforcement Learning via Approximate Sampling.* RLJ.

[190] Ric De Santi, Manish Prajapat, and Andreas Krause (2024). *Global Reinforcement Learning: Beyond Linear and Convex Rewards via Submodular Semi-gradient Methods.* International Conference on Machine Learning.

[191] Van-Hau Pham, Do Thi Thu Hien, Nguyen Phuc Chuong, et al. (2024). *A Coverage-Guided Fuzzing Method for Automatic Software Vulnerability Detection Using Reinforcement Learning-Enabled Multi-Level Input Mutation.* IEEE Access.

[192] Wei Ding, Siyang Jiang, Hsi-Wen Chen, et al. (2023). *Incremental Reinforcement Learning with Dual-Adaptive -Greedy Exploration.* AAAI Conference on Artificial Intelligence.

[193] Jesu Narkarunai Arasu Malaiyappan, Sai Mani Krishna Sistla, and Jawaharbabu Jeyaraman (2024). *Advancements in Reinforcement Learning Algorithms for Autonomous Systems*. International Journal of Innovative Science and Research Technology.

[194] Xuemei Gan, Ying Zuo, Ansi Zhang, et al. (2023). *Digital twin-enabled adaptive scheduling strategy based on deep reinforcement learning*. Science China Technological Sciences.

[195] Kai-Wen Zhao, Yi Ma, Jinyi Liu, et al. (2023). *Ensemble-based Offline-to-Online Reinforcement Learning: From Pessimistic Learning to Optimistic Exploration*. arXiv.org.

[196] Kaidi Xu, Shenglong Zhou, and Geoffrey Ye Li (2023). *Federated Reinforcement Learning for Resource Allocation in V2X Networks*. IEEE Vehicular Technology Conference.

[197] Nesrine Khlif, Khraief-Hadded Nahla, and Belghith Safya (2023). *Reinforcement learning with modified exploration strategy for mobile robot path planning*. Robotica (Cambridge. Print).

[198] S. Sreedharan, and Michael Katz (2023). *Optimistic Exploration in Reinforcement Learning Using Symbolic Model Estimates*. Neural Information Processing Systems.

[199] Siyuan Guo, Yanchao Sun, Jifeng Hu, et al. (2023). *A Simple Unified Uncertainty-Guided Framework for Offline-to-Online Reinforcement Learning*. arXiv.org.

[200] Xiao Zhang, Hai Zhang, Hongtu Zhou, et al. (2023). *Safe Reinforcement Learning With Dead-Ends Avoidance and Recovery*. IEEE Robotics and Automation Letters.

[201] Ali Beikmohammadi, and S. Magnússon (2023). *TA-Explore: Teacher-Assisted Exploration for Facilitating Fast Reinforcement Learning*. Adaptive Agents and Multi-Agent Systems.

[202] Qisong Yang, T. D. Simão, N. Jansen, et al. (2023). *Reinforcement Learning by Guided Safe Exploration*. European Conference on Artificial Intelligence.

[203] Jonatan Alvarez, Assia Belbachir, Faiza Belbachir, et al. (2023). *Forest Fire Localization: From Reinforcement Learning Exploration to a Dynamic Drone Control*. Journal of Intelligent and Robotic Systems.

[204] Tianyi Li, Gen-ke Yang, and Jian Chu (2023). *Implicit Posteriori Parameter Distribution Optimization in Reinforcement Learning*. IEEE Transactions on Cybernetics.

[205] Yuan Zi, Lei Fan, Xuqing Wu, et al. (2023). *Active Gamma-Ray Log Pattern Localization With Distributionally Robust Reinforcement Learning*. IEEE Transactions on Geoscience and Remote Sensing.

[206] Junjun Yang, Kaige Tan, Lei Feng, et al. (2023). *Reducing the Learning Time of Reinforcement Learning for the Supervisory Control of Discrete Event Systems*. IEEE Access.

[207] Chuxiong Sun, Rui Wang, Qian Li, et al. (2021). *Reward Space Noise for Exploration in Deep Reinforcement Learning*. International journal of pattern recognition and artificial intelligence.

[208] Zijing Guo, Chendie Yao, Yanghe Feng, et al. (2022). *Survey of Reinforcement Learning based on Human Prior Knowledge*. Journal of Uncertain Systems.

[209] Sahisnu Mazumder, Bing Liu, Shuai Wang, et al. (2022). *Knowledge-Guided Exploration in Deep Reinforcement Learning*. arXiv.org.

[210] Ji-Yun Oh, Joonkee Kim, and Se-Young Yun (2022). *Risk Perspective Exploration in Distributional Reinforcement Learning*. arXiv.org.

[211] J. Vidaković, B. Jerbić, B. Šekoranja, et al. (2020). *Accelerating Robot Trajectory Learning for Stochastic Tasks*. IEEE Access.

[212] Xianzhuo Sun, Zhao Xu, Jing Qiu, et al. (2024). *Optimal Volt/Var Control for Unbalanced Distribution Networks With Human-in-the-Loop Deep Reinforcement Learning*. IEEE Transactions on Smart Grid.

[213] Peipei Yu, Zhen-yu Wang, Hongcai Zhang, et al. (2024). *Safe Reinforcement Learning for Power System Control: A Review*. arXiv.org.

[214] Ming Wang, Jie Zhang, Peng Zhang, et al. (2024). *Cooperative multi-agent reinforcement learning for multi-area integrated scheduling in wafer fabs*. International Journal of Production Research.

[215] Yunlong Ding, Minchi Kuang, Heng Shi, et al. (2024). *Multi-UAV Cooperative Target Assignment Method Based on Reinforcement Learning*. Drones.

[216] Jingwen Yang, Ping Wang, and Yongfeng Ju (2024). *Variable Speed Limit Intelligent Decision-Making Control Strategy Based on Deep Reinforcement Learning under Emergencies*. Sustainability.

[217] Sajjad Afroosheh, Khodakhast Esapour, Reza Khorram-Nia, et al. (2024). *Reinforcement learning layout-based optimal energy management in smart home: AI-based approach*. IET Generation, Transmission amp; Distribution.

[218] Yihong Dong, Xue Jiang, Yongding Tao, et al. (2025). *RL-PLUS: Countering Capability Boundary Collapse of LLMs in Reinforcement Learning with Hybrid-policy Optimization*. arXiv.org.

[219] Qingling Zhu, Xiaoqiang Wu, Qiuzhen Lin, et al. (2024). *Two-Stage Evolutionary Reinforcement Learning for Enhancing Exploration and Exploitation*. AAAI Conference on Artificial Intelligence.

[220] Xuanchen Xiang, Ruisheng Diao, S. Bernadin, et al. (2024). *An Intelligent Parameter Identification Method of DFIG Systems Using Hybrid Particle Swarm Optimization and Reinforcement Learning*. IEEE Access.

[221] Ji Qi, Haibo Gao, Huanli Su, et al. (2024). *Reinforcement Learning and Sim-to-Real Transfer of Reorientation and Landing Control for Quadruped Robots on Asteroids.* IEEE transactions on industrial electronics (1982. Print).

[222] Bolei Zhang, Fu Xiao, and Lifa Wu (2024). *Offline Reinforcement Learning for Asynchronous Task Offloading in Mobile Edge Computing.* IEEE Transactions on Network and Service Management.

[223] Siqing Sun, Huachao Dong, and Tianbo Li (2024). *A modified evolutionary reinforcement learning for multi-agent region protection with fewer defenders.* Complex amp; Intelligent Systems.

[224] Dipo Dunsin, Mohamed Chahine Ghanem, Karim Ouazzane, et al. (2024). *Reinforcement Learning for an Efficient and Effective Malware Investigation during Cyber Incident Response.* arXiv.org.

[225] Haotian Hu, Yiqin Yang, Jianing Ye, et al. (2024). *Bayesian Design Principles for Offline-to-Online Reinforcement Learning.* International Conference on Machine Learning.

[226] Chang-Hoon Ji, Dong-Hee Shin, Young-Han Son, et al. (2024). *Sparse Graph Representation Learning Based on Reinforcement Learning for Personalized Mild Cognitive Impairment (MCI) Diagnosis.* IEEE journal of biomedical and health informatics.

[227] Simone Parisi, Alireza Kazemipour, and Michael Bowling (2024). *Beyond Optimism: Exploration With Partially Observable Rewards.* Neural Information Processing Systems.

[228] Yiming Wang, Kaiyan Zhao, Furui Liu, et al. (2024). *Rethinking Exploration in Reinforcement Learning with Effective Metric-Based Exploration Bonus.* Neural Information Processing Systems.

[229] Fan Wu, Rui Zhang, Qi Yi, et al. (2024). *OCEAN-MBRL: Offline Conservative*

Exploration for Model-Based Offline Reinforcement Learning. AAAI Conference on Artificial Intelligence.

[230] Dongfang Zhao, Huanshi Xu, and Zhang Xun (2024). *Active Exploration Deep Reinforcement Learning for Continuous Action Space with Forward Prediction*. International Journal of Computational Intelligence Systems.

[231] Hean Hua, Yaonan Wang, Hang Zhong, et al. (2025). *A Novel Guided Deep Reinforcement Learning Tracking Control Strategy for Multirotors*. IEEE Transactions on Automation Science and Engineering.

[232] Runpeng Dai, Linfeng Song, Haolin Liu, et al. (2025). *CDE: Curiosity-Driven Exploration for Efficient Reinforcement Learning in Large Language Models*. Unpublished manuscript.

[233] Xin Chang, Yanbin Li, Guanjie Zhang, et al. (2024). *An Improved Reinforcement Learning Method Based on Unsupervised Learning*. IEEE Access.

[234] J. Janjua, Shagufta Kousar, Areeba Khan, et al. (2024). *Enhancing Scalability in Reinforcement Learning for Open Spaces*. 2024 International Conference on Decision Aid Sciences and Applications (DASA).

[235] Jorge Val Ledesma, Rafał Wiśniewski, C. Kallesøe, et al. (2024). *Water Age Control for Water Distribution Networks via Safe Reinforcement Learning*. IEEE Transactions on Control Systems Technology.

[236] Mingkang Wu, Umer Siddique, Abhinav Sinha, et al. (2024). *Offline Reinforcement Learning with Failure Under Sparse Reward Environments*. International Conference on Multimodal Interaction.

[237] Homayoun Honari, Amir M. Soufi Enayati, Mehran Ghafarian Tamizi, et al. (2024). *Meta SAC-Lag: Towards Deployable Safe Reinforcement Learning via MetaGradient-based Hyperparameter Tuning*. IEEE/RJS International Conference on Intelligent RObots and Systems.

[238] Jiamin Shi, Tangyike Zhang, Ziqi Zong, et al. (2024). *Task-Driven Autonomous Driving: Balanced Strategies Integrating Curriculum Reinforcement Learning and Residual Policy*. IEEE Robotics and Automation Letters.

[239] Thinh Lu, Divyam Sobti, Deepak Talwar, et al. (2025). *Reinforcement learning-based dynamic field exploration and reconstruction using multi-robot systems for environmental monitoring*. Frontiers Robotics AI.

[240] Muhan Hou, Koen V. Hindriks, Guszti Eiben, et al. (2024). *"Give Me an Example Like This": Episodic Active Reinforcement Learning from Demonstrations*. International Conference on Human-Agent Interaction.