# A Survey of Reinforcement Learning and Its Modern Variants




## Abstract

Reinforcement Learning (RL) is a dynamic machine learning paradigm that empowers agents to make sequential decisions by interacting with their environment to optimize cumulative rewards. This survey paper offers a comprehensive examination of RL, emphasizing its modern variants, including Offline RL, Decision Transformer (DT), Deep RL, policy optimization, and model-free methods. Offline RL, which learns from static datasets without further environment interaction, addresses challenges like high variance and distribution shifts, offering advantages in safety-critical domains. The Decision Transformer redefines RL by framing it as a sequence modeling task, leveraging transformer architectures for enhanced decision-making. Deep RL integrates deep learning with RL to handle high-dimensional spaces, though it faces challenges like sample inefficiency and exploration-exploitation trade-offs. Policy optimization techniques focus on refining policies for improved performance, with distinctions between on-policy and off-policy methods. Model-free methods learn policies directly from environmental interactions, providing robustness in complex scenarios. The survey underscores RL's applicability across domains such as robotics, finance, and healthcare, while addressing ethical considerations like interpretability and safety. Future directions include enhancing sample efficiency, integrating control theory, and exploring hierarchical and multi-objective approaches. This survey highlights RL's potential to revolutionize decision-making across various sectors, advocating for continued research to overcome existing challenges and fully realize its capabilities.


## 1 Introduction

### 1.1 Understanding Reinforcement Learning

Reinforcement Learning (RL) is a fundamental machine learning paradigm that enables agents to make decisions in environments to maximize cumulative rewards. Central to RL is sequential decision-making, where agents learn optimal policies through interactions with their environment, refining their actions based on feedback received as rewards [1]. This learning process requires a delicate balance between exploration and exploitation, where agents must choose between exploring new actions or exploiting known rewarding ones [2].

The significance of RL is evident in its application to complex decision-making scenarios, particularly in environments with partially observable states [3]. In high-stakes fields such as healthcare and finance, RL's traditional focus on maximizing expected outcomes without accounting for associated risks highlights the necessity for risk-sensitive approaches [4]. Moreover, RL's relevance in agent-based computational economics (ACE) underscores its potential to address economic policy challenges [5].

RL agents frequently encounter challenges in environments characterized by sparse rewards and large decision spaces, necessitating innovative strategies to enhance decision-making efficacy [6]. The

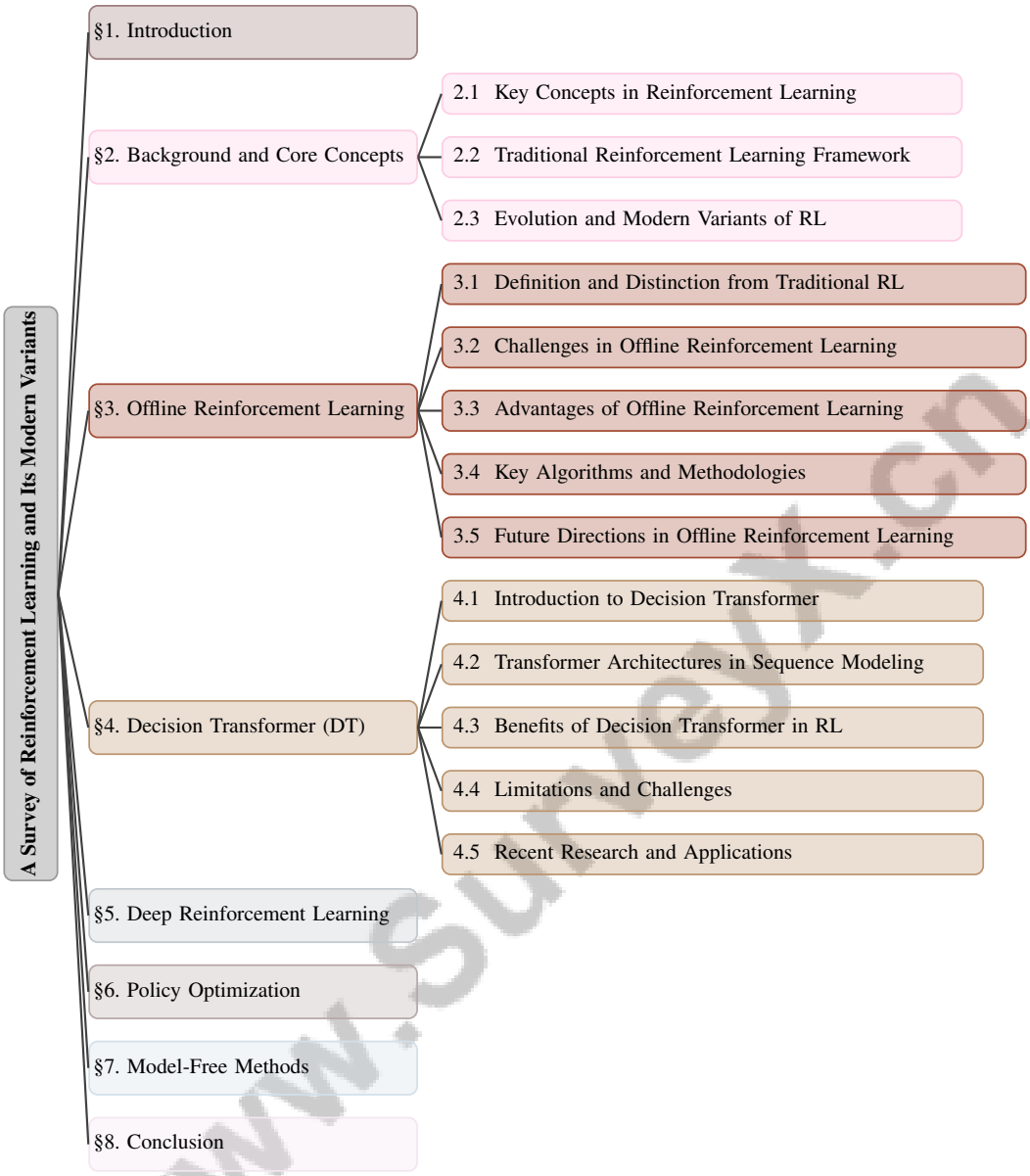


Figure 1: chapter structure

timing of actions is also critical, as RL emphasizes not only how to act but when to act, which is vital for optimizing decision strategies [7].

Despite its advancements, RL continues to evolve, addressing challenges such as delayed rewards, where actions leading to successful outcomes may not receive immediate positive feedback, complicating the reinforcement of beneficial behaviors [1]. As RL develops, its adaptability and potential for synthesizing control policies for diverse applications further underscore its importance in contemporary machine learning.

### 1.2 Motivation and Significance

This survey on Reinforcement Learning (RL) is motivated by its transformative potential across various domains, particularly in enhancing decision-making processes under uncertainty [8]. However, the real-world application of RL is often limited by safety concerns. This survey addresses these issues by exploring advancements that integrate risk-awareness and safety considerations into RL

frameworks, which is crucial in safety-critical applications like autonomous driving, where precision in decision-making is essential.

Additionally, the survey aims to bridge significant gaps in the literature by examining the integration of Knowledge Representation and Reasoning (KRR) methods to enhance RL's efficiency, generalization, and interpretability [9]. This integration is particularly relevant for complex real-life problems where traditional RL approaches may be inadequate. The survey also emphasizes the significance of Automated Reinforcement Learning (AutoRL) by providing a unified taxonomy and identifying open problems, thereby fostering further research in automating the RL process [10].

In dynamic environments such as large-scale network resource management, RL's capability to optimize resource allocation highlights its importance in achieving efficient and adaptive solutions [11]. Furthermore, RL's application in optimizing customer relationships and maximizing user lifetime value (LTV) exemplifies its impact on real-world challenges [3]. The survey also explores RL's role in Cloud autoscaling, addressing the need for dynamic resource management policies that can optimize application execution in stochastic environments [12].

The necessity of developing explainable and interpretable RL models is crucial for user acceptance and system efficiency in applications like home energy management systems (HEMS) and HVAC systems in smart buildings [7]. Moreover, the survey documents the ethical considerations associated with artificial RL agents, emphasizing the importance of responsible deployment in decision-making tasks [8].

By exploring these diverse aspects, this survey provides a comprehensive overview of RL's foundational concepts, challenges, and applications, underscoring the need for continued research and innovation to fully realize its potential in various sectors, including economic policy modeling, where recent developments in RL may overcome the historical limitations of classical agent-based techniques [5].

### 1.3 Overview of the Survey Structure

This survey is structured to offer a thorough examination of Reinforcement Learning (RL) and its modern variants, focusing on foundational principles and recent advancements. Initial sections introduce core RL concepts, including agents, environments, states, actions, and rewards, while discussing the traditional RL framework and its objectives. The survey then transitions to the evolution of RL into modern variants, setting the stage for detailed discussions in subsequent sections.

In-depth analysis of Offline Reinforcement Learning follows, highlighting its definition, challenges, advantages, and key algorithms. The subsequent section provides a comprehensive examination of the Decision Transformer (DT) model, emphasizing its innovative adaptation of transformer architectures for sequence modeling within offline RL tasks. This analysis discusses the model's advantages, such as leveraging pre-collected datasets and enhancing performance through prompt-based techniques, alongside its limitations regarding stitching capabilities and data requirements. Recent applications of the DT and its variants, including the Q-learning Decision Transformer and the Bootstrapped Transformer, are explored, showcasing their contributions to overcoming challenges in offline RL and improving policy learning from sub-optimal trajectories [13, 14, 15].

Further sections delve into Deep Reinforcement Learning, exploring its integration with deep neural networks to manage high-dimensional state and action spaces, and highlighting notable algorithms and applications. The discussion then shifts to Policy Optimization, explaining techniques for directly improving policies and examining the differences between on-policy and off-policy methods.

The survey also covers Model-Free Methods, discussing their significance in learning policies or value functions without constructing an explicit model of the environment, while comparing them with model-based methods. Concluding sections reflect on the current state of RL and its modern variants, emphasizing ethical and practical considerations, challenges, applications, and implications across various domains.

Throughout the survey, the historical progression from foundational algorithms like Q-learning to state-of-the-art methods such as TD3 and PPO is explored, while specific applications are intentionally excluded to maintain focus on the algorithms themselves [16].The following sections are organized as shown in Figure 1.

## 2 Background and Core Concepts

### 2.1 Key Concepts in Reinforcement Learning

Reinforcement Learning (RL) is centered on the interaction between agents and their environments to develop policies that optimize cumulative rewards [3]. This interaction is characterized by sequences of states, actions, and rewards, with states representing the current context and policies mapping states to actions to maximize expected rewards [16, 17]. RL is mathematically structured as a Markov Decision Process (MDP), which models decision-making under uncertainty, influenced by stochastic elements and the agent's actions [7]. A major challenge in RL is the exploration-exploitation tradeoff, where agents must balance exploring new actions to gather information against exploiting known actions to gain rewards [18]. This balance is crucial in complex environments with high-dimensional state and action spaces, which can hinder learning efficiency [19].

Model-free RL methods, which derive policies directly from interactions with the environment without constructing explicit models, demonstrate the practical application of RL concepts [16]. These methods, however, often suffer from sample inefficiency, necessitating numerous trials to develop effective policies [20]. Recent advances in RL aim to enhance exploration strategies and data efficiency through intrinsic motivation and causal principles, addressing traditional exploration challenges [21]. Understanding these core concepts is essential for constructing effective RL frameworks applicable in various domains, from optimizing large-scale network resource allocation to managing customer relationships in partially observable settings. Integrating human knowledge into RL, as seen in Knowledge-Based Reinforcement Learning (KB-RL), enhances decision-making in complex environments [22]. RL's application in Cloud autoscaling categorizes approaches into model-based and model-free techniques, reflecting the ongoing evolution of RL methodologies to tackle diverse challenges [12].

### 2.2 Traditional Reinforcement Learning Framework

The traditional Reinforcement Learning (RL) framework is fundamentally based on the Markov Decision Process (MDP) and models decision-making in stochastic environments [16]. This framework involves an agent interacting with its environment through states, actions, and rewards, aiming to derive an optimal policy that maximizes cumulative rewards. The core components—agents, environments, states, actions, and rewards—define RL's operational dynamics [16]. A central challenge in traditional RL is the exploration-exploitation tradeoff, requiring agents to balance exploring new actions for information against exploiting known actions for immediate rewards [18]. This dilemma is exacerbated in high-dimensional and continuous control tasks, where existing methods often struggle without extensive reward shaping or environment tuning [18]. Furthermore, the inefficiency of fixed state representations in temporal difference (TD) learning limits value prediction accuracy, complicating the learning process [23].

Traditional RL methods often prioritize action optimization without considering timing, potentially leading to suboptimal performance in dynamic environments [7]. The lack of scalable algorithms adhering to generalized policy iteration (GPI) and trust-region learning (TRL) frameworks complicates optimization, frequently relying on empirical heuristics without guaranteed performance [24]. Additionally, traditional RL faces risks associated with unsafe actions during learning, jeopardizing system integrity, such as in power converters [25]. Static reward functions may fail to capture real-world dynamics, highlighting the need for risk-aware decision-making strategies in average-reward MDPs [26].

Despite these challenges, the traditional RL framework remains foundational, serving as a basis for evaluating new methodologies and algorithms. Efforts to address its limitations include leveraging duality theory to enhance optimization and developing methods to mitigate feature correlation in deep reinforcement learning [27]. Integrating human knowledge into RL frameworks, as seen in KB-RL, also aims to tackle issues like data dependency and explainability in Neural Networks (NN) [22]. These advancements are crucial for improving RL's applicability across diverse domains. Recent innovations have sought to enhance sample utilization and convergence speed in on-policy RL algorithms, particularly in challenging continuous robotic control tasks [28]. Constructing a set of policies that can be combined to yield optimal solutions for new tasks with linearly-expressed reward functions has also been explored to improve adaptability within traditional RL frameworks

[29]. These innovations are vital for overcoming inherent limitations and broadening the applicability of traditional RL methods to complex real-world problems.

### 2.3 Evolution and Modern Variants of RL

The evolution of Reinforcement Learning (RL) into modern variants has been marked by advancements enhancing scalability, efficiency, and adaptability to complex environments. A pivotal development is the integration of deep learning architectures, significantly improving RL's capacity to process high-dimensional data and extract intricate patterns for effective policy learning [22]. This is exemplified by models like the Trajectory Transformer, which utilizes sequence modeling to predict sequences of states, actions, and rewards, thereby enhancing adaptability and efficiency [30].

The introduction of Knowledge-Based Reinforcement Learning (KB-RL) marks a significant milestone, combining knowledge-based systems with RL to optimize decision-making in complex scenarios, such as strategy games [22]. This approach leverages existing domain knowledge to improve learning efficiency and decision-making accuracy, addressing traditional challenges faced by RL in dynamic environments. Recent innovations in policy learning have emerged from hybrid architectures that merge offline policies trained on expert demonstrations with online learning strategies, such as Proximal Policy Optimization (PPO). These architectures often incorporate advanced reward shaping mechanisms to accelerate policy optimization [28]. The application of Hamilton-Jacobi-Bellman (HJB) equations and first-order gradients in learning time-varying value and Lyapunov functions exemplifies the shift towards sophisticated control strategies integral to modern RL frameworks [31].

The exploration of meta-gradient approaches for discovering options and metalearned subgoals further underscores the evolution of RL, enhancing learning flexibility and efficiency [32]. Applying duality theory in RL, inspired by its success in supervised learning, offers potential benefits in optimizing value functions and policy learning, providing a more robust foundation for RL algorithms [33]. In offline RL, empirical studies have shown that replacing regression with classification for value function estimation can improve performance across various tasks, highlighting the potential for methodological innovations to enhance RL applicability [34]. Methods like Discrete-to-Deep Supervised Policy Learning (D2D-SPL) aim to discretize continuous state spaces and utilize actor-critic frameworks, reflecting ongoing efforts to refine RL methodologies [20].

The introduction of the Mirror Learning framework exemplifies RL's evolution by bridging the gap between theoretical guarantees and practical performance, providing a robust foundation for RL algorithms [24]. The development of Distributional Successor Features for Zero-Shot Policy Optimization (DiSPOs) aims to enhance transferability across reward functions while mitigating compounding errors, further illustrating RL's dynamic evolution [35]. Modern RL variants also benefit from advancements in representation learning, with methods like Disentangled Environment and Agent Representations (DEAR) improving sample efficiency in visual RL through distinct representations of the environment and agent [21]. Theoretical characterizations of state representations learned by TD learning have revealed differences from those learned by Monte Carlo and residual gradient algorithms, offering insights into optimizing representation dynamics.

These advancements collectively underscore the dynamic nature of RL's evolution, emphasizing its expanding role in addressing complex real-world challenges across various domains, from network resource management to autonomous systems [11]. As RL continues to evolve, its modern variants are poised to deliver increasingly sophisticated solutions to the multifaceted demands of contemporary decision-making tasks, including complex lexicographic multi-objective reinforcement learning (MORL) problems requiring subtask prioritization [36].

In recent years, Offline Reinforcement Learning (RL) has garnered significant attention within the field due to its unique advantages and challenges. As depicted in Figure 2, this figure illustrates the hierarchical structure of Offline RL, detailing its definition, challenges, advantages, key algorithms, and future directions. The figure effectively highlights the distinction between Offline RL and traditional RL, emphasizing the specific challenges posed by dataset limitations and exploration constraints. Furthermore, it underscores the advantages of Offline RL, particularly in terms of safety and cost efficiency. The methodologies that enhance its robustness and data efficiency are also presented, providing a comprehensive overview of the current landscape. Lastly, the future directions outlined in the figure emphasize the potential for algorithmic enhancements and hybrid approaches, which promise to broaden the applicability of Offline RL across various domains.

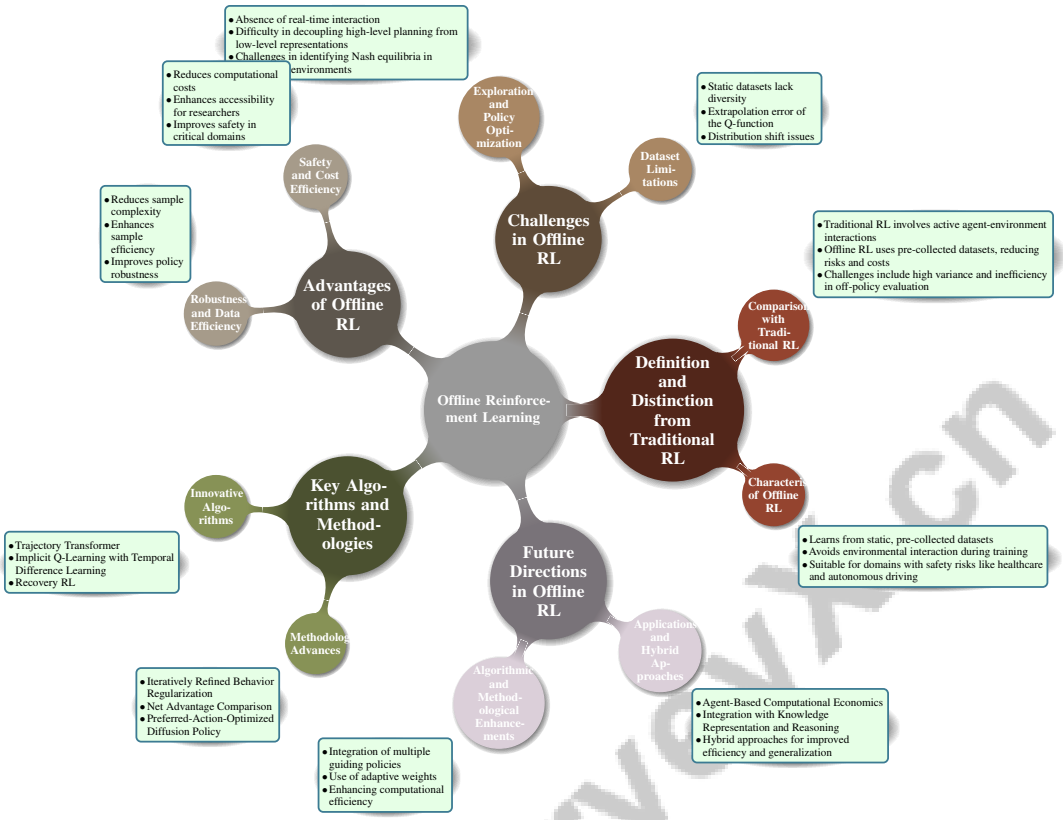


Figure 2: This figure illustrates the hierarchical structure of Offline Reinforcement Learning, detailing its definition, challenges, advantages, key algorithms, and future directions. It highlights the distinction between Offline RL and traditional RL, the challenges faced due to dataset limitations and exploration constraints, the advantages in terms of safety and cost efficiency, and the methodologies that enhance its robustness and data efficiency. Future directions emphasize algorithmic enhancements and hybrid approaches for broader applicability.

# 3 Offline Reinforcement Learning

## 3.1 Definition and Distinction from Traditional RL

Offline Reinforcement Learning (Offline RL) is characterized by learning optimal policies from static, pre-collected datasets, avoiding further environmental interaction during training [37]. This approach is particularly advantageous in domains like healthcare and autonomous driving, where real-time exploration poses safety risks [38]. Offline RL mitigates the dangers and costs of real-time data collection, making it suitable for scenarios where exploration is hazardous [12].

The primary distinction between Offline and traditional RL lies in data collection and policy optimization. Traditional RL involves active agent-environment interactions for real-time feedback to iteratively refine policies, which can be inefficient and risky, especially in non-stationary environments [39]. Offline RL, on the other hand, uses pre-collected datasets, reducing risks and costs associated with real-time interaction, particularly where exploration could lead to unsafe outcomes [38].

Offline RL faces challenges such as high variance and inefficiency in off-policy evaluation, which can hinder iterative algorithms [40]. The absence of online corrective feedback necessitates robust methodologies to ensure learned policies generalize well to unseen situations, addressing distributional shifts between training datasets and real-world scenarios [41]. Furthermore, existing methodologies often struggle to effectively utilize high-performing trajectories in mixed datasets dominated by low-performing ones [39].

Despite these challenges, Offline RL offers opportunities to enhance learning efficiency and policy robustness by leveraging large, diverse datasets. Techniques like Risk-Aware Reward Shaping (RARS) incorporate risk awareness into the reward function, improving policy performance in risk-sensitive applications [36]. Additionally, learning goal-conditioned policies from offline data, especially in scenarios with sparse rewards and complex relabeling strategies, remains an active research area [22]. The SFOLS algorithm, for example, iteratively learns a set of policies whose successor features form a convex coverage set (CCS), facilitating optimal policy transfer [29]. Furthermore, determining optimal policy transitions within historical data constraints is a critical focus [42].

As Offline RL evolves, it promises to broaden the scope of RL applications by enabling safe and efficient learning in constrained environments, paving the way for broader adoption in sectors where traditional RL methods fall short [37]. The limitations of existing offline RL methods that utilize weighted regression, which restrict policy improvement to dataset-collected actions and struggle with the out-of-distribution (OOD) problem, underscore the need for innovative approaches to overcome these challenges [43].

### 3.2 Challenges in Offline Reinforcement Learning

Offline Reinforcement Learning (RL) faces significant challenges due to its reliance on static datasets, which often lack the diversity needed for robust policy learning. A primary challenge is the extrapolation error of the Q-function, leading to performance degradation when transitioning between environments, despite retaining all past data [44]. This issue is exacerbated by the static nature of offline datasets, which ties algorithms to the behavior policy from which the dataset was collected, underutilizing high-return trajectories [45].

The problem of distribution shift is particularly acute, as offline RL methods struggle to generalize policies to unseen states due to constraints like concentrability and linear q-realizability, requiring a sample size that scales with the number of states in the Markov Decision Process (MDP) [46]. Additionally, methods like Goal-Conditioned Supervised Learning (GCSL) often produce suboptimal policies by uniformly weighting experiences, failing to leverage the importance of different relabeling goals [47].

A significant challenge in offline RL is the absence of real-time interaction with the environment, complicating policy optimization as offline RL lacks exploratory mechanisms necessary for low sub-optimality when extrapolating from historical data [37]. This lack of exploration is further compounded by the conflict between exploring new tasks and limiting exploration to avoid constraint violations, potentially resulting in suboptimal policy optimization and unsafe actions [38].

Offline RL also faces difficulties in decoupling high-level planning from low-level representations, limiting the effectiveness of existing methods due to the complexity of integrating multimodal data [48]. Moreover, determining the root cause of poor performance—whether due to an imperfect value function, inadequate policy extraction methods, or insufficient generalization—remains a critical issue [49].

The lack of interaction with opponents in competitive environments further complicates the identification of Nash equilibria, as offline RL struggles with incomplete state-action coverage in real-world datasets [50]. Existing methods are constrained by the collected data, often representing only a small portion of the entire state-action space, leading to limited policy performance and instability due to Q-value estimation errors [43]. Accurately modeling switching costs, which can vary significantly based on the policies involved and the context, poses additional challenges in deciding whether a switch is beneficial [42]. The inherent uncertainty and potential errors in policy estimation when deviating from any guiding policy can lead to suboptimal performance [51].

The challenges in offline reinforcement learning (RL) underscore the urgent need for innovative strategies that effectively integrate offline data utilization with real-world applications. This integration is crucial for developing offline RL methodologies that can navigate dynamic environments, leveraging insights from recent studies that highlight the importance of policy extraction, generalization, and dataset quality in achieving optimal performance. Addressing these factors can enhance offline RL's effectiveness, ensuring it meets the demands of practical scenarios where interaction with the environment is limited [49, 52, 53, 54, 55].

### 3.3 Advantages of Offline Reinforcement Learning

Offline Reinforcement Learning (Offline RL) offers significant advantages, particularly in scenarios where real-time data collection is impractical or risky. It leverages pre-collected datasets, reducing computational costs and enhancing accessibility for researchers without extensive resources, facilitating faster development and broader applicability [56]. This approach is advantageous in safety-critical domains like autonomous driving, where ensuring compliance with constraints is paramount. Techniques such as Risk-Aware Reward Shaping (RARS) enhance safety and exploration capabilities, leading to improved performance [36].

Offline RL methods reduce sample complexity by using offline data to accelerate convergence to optimal policies. Approaches like QCSE enhance sample efficiency by addressing distribution shifts and promoting exploration [57]. Adversarial training methods, such as ARMOR, ensure learned policies do not perform worse than reference policies, enhancing policy robustness [40].

Offline RL's robustness across various environments is further highlighted by methods like the model-free two-step design, which improves transient learning performance and reduces wear on systems during learning, enhancing its applicability in industrial settings [28]. Methods such as the Preferred-Action-Optimized Diffusion Policy (PAO-DP) optimize diffusion policies using preferred actions generated through a critic function, addressing the out-of-distribution (OOD) problem [43].

In terms of data efficiency, Offline RL benefits from approaches like Weighted Goal-Conditioned Supervised Learning (WGCSL), which effectively handle sparse rewards, maintain stability, and perform well across various tasks [47]. The integration of structured, temporally abstract guidance in the form of intent embeddings, as demonstrated by IQL-TD-MPC, further enhances the performance of offline RL algorithms [39].

The advantages of Offline Reinforcement Learning (RL) are substantial, offering reduced computational costs, enhanced safety, and improved sample efficiency. These benefits arise from its ability to leverage existing knowledge for policy optimization, making it a powerful tool for diverse applications. Offline RL can outperform traditional imitation learning even with lower quality data by utilizing value functions effectively. Recent advancements, such as on-policy Q estimates for policy improvement and better sample reuse techniques, have demonstrated significant performance gains and robustness across various environments. Frameworks like Expert-Supervised RL enhance interpretability and safety by incorporating uncertainty quantification, allowing for tailored risk management in practical applications. Collectively, these features position Offline RL as an essential approach in the evolving landscape of machine learning [58, 56, 49, 59, 60]. These benefits are particularly pronounced in domains where traditional RL methods are constrained by the need for real-time interaction and exploration, making Offline RL a compelling alternative for complex decision-making tasks.

### 3.4 Key Algorithms and Methodologies

Offline Reinforcement Learning (RL) has advanced significantly through innovative algorithms and methodologies optimizing policy learning from static datasets. The Trajectory Transformer treats states, actions, and rewards as a single data sequence, capturing long-horizon dependencies and enhancing adaptability [41]. Implicit Q-Learning with Temporal Difference Learning for Model Predictive Control (IQL-TD-MPC) integrates model-based RL with offline learning to improve planning and decision-making in sparse-reward tasks [39].

The Recovery RL algorithm employs two distinct policies to manage task performance and safety, using offline data to enhance learning about potential constraint violations. It combines offline learning of recovery zones with real-time policy execution, ensuring safe exploration and effective task learning [38]. The Iteratively Refined Behavior Regularization (CPI) algorithm refines the reference policy for behavior regularization while maintaining updates within the reference policy's support, ensuring stability [40].

The Net Advantage Comparison (NAC) algorithm balances potential gains from policy switching with associated costs, using net values and Q-functions to inform decisions. This involves evaluating the old policy, alternately improving and evaluating the new policy, and comparing their net values to decide on switching [42]. Additionally, the Guided Reinforcement Learning with Monte Carlo

Tree Search (GRL-MCTS) method incorporates MCTS as a guiding policy to enhance RL agents' learning and performance [51].

The Preferred-Action-Optimized Diffusion Policy (PAO-DP) optimizes a diffusion-based policy using preferred actions derived from a critic function, enhancing performance in offline RL [43]. The CUOLR method treats the ranking process as a Markov Decision Process (MDP), allowing offline RL algorithms to optimize ranking lists [61].

These algorithms and methodologies collectively advance offline RL's capabilities, providing robust solutions for scenarios where real-time interaction is limited or impractical. Recent innovations in offline reinforcement learning (RL) are broadening its applicability across diverse sectors by integrating offline datasets and online interactions. This fusion enhances decision-making efficiency in complex environments, as demonstrated by advancements like the FineTuneRL setting, which optimally utilizes offline data to reduce the need for online interactions, and methods harmonizing pre-trained offline policies with online exploration strategies. New algorithms have been introduced to facilitate learning multiple distinct solutions from a single task in offline RL, enriching the toolkit for tackling intricate decision-making challenges [62, 53, 63].

As illustrated in Figure 3, this figure categorizes key algorithms in offline reinforcement learning into trajectory optimization, safety and stability, and policy improvement. Each category includes specific algorithms that contribute to the enhancement of offline RL capabilities, demonstrating advancements in handling complex decision-making tasks without real-time interaction. The first subfigure presents a comparative analysis of Elo ratings across various strategies in a neural network evaluation task, highlighting the performance of Online ExIt (exponential), Online ExIt (buffer), Batch ExIt, and REINFORCE strategies. This graph serves as a crucial tool for understanding how different strategies perform over numerous neural network evaluations, providing a quantitative measure of their effectiveness. The second subfigure presents a scatter plot examining the SOLO performance against itself to reveal potential correlations or patterns. Although the scatter plot shows data points without a distinct trend, it underscores the complexity and variability inherent in reinforcement learning tasks. Together, these visualizations encapsulate the diversity of approaches and the intricate dynamics involved in offline reinforcement learning, setting the stage for further exploration and refinement of these methodologies [64, 65].

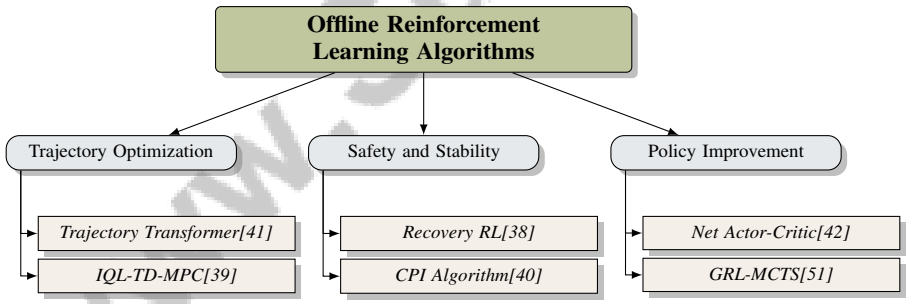


Figure 3: This figure illustrates key algorithms in offline reinforcement learning, categorized into trajectory optimization, safety and stability, and policy improvement. Each category includes specific algorithms that contribute to the enhancement of offline RL capabilities, demonstrating advancements in handling complex decision-making tasks without real-time interaction.

### 3.5 Future Directions in Offline Reinforcement Learning

The trajectory of Offline Reinforcement Learning (RL) research is set to focus on several key areas that promise to enhance its applicability and effectiveness across various domains. One promising direction is the integration of multiple guiding policies and the use of adaptive weights to modulate their influence, significantly enhancing the adaptability and robustness of RL algorithms in complex environments [51]. This approach could be particularly beneficial in domains characterized by sparse rewards and intricate task structures, as exemplified by D4RL benchmark environments like Kitchen and AntMaze [43].

There is also a compelling need to advance the computational efficiency of RL algorithms, especially in high-dimensional and continuous action spaces. Future research could explore the potential of

Transformer architectures and dynamic programming techniques to enhance planning capabilities and improve handling of continuous inputs [41]. Additionally, refining cost modeling and applying algorithms such as Net Advantage Comparison (NAC) in diverse real-world scenarios could bolster decision-making processes [42].

In economic policy, deeper analyses of RL methods within Agent-Based Computational Economics (ACE) are warranted. This involves comparing different RL approaches and exploring larger-scale projects leveraging modern RL techniques to address complex economic challenges [5]. Such endeavors could provide valuable insights into the scalability and adaptability of RL methodologies in dynamic economic environments.

Exploring hybrid approaches that integrate RL with other methodologies, such as Knowledge Representation and Reasoning (KRR), holds promise for improving RL's efficiency, generalization, and interpretability [37]. This integration could be pivotal in enhancing RL's applicability in domains requiring sophisticated decision-making capabilities, such as cloud autoscaling and other resource management tasks [12].

The rapidly advancing field of offline reinforcement learning (RL) research highlights the promise of innovative methodologies, such as the Bootstrapped Transformer algorithm and the FineTuneRL setting, which aim to enhance policy learning from static trajectory data and optimize the use of offline datasets in conjunction with online interactions. These approaches address critical limitations in existing offline RL training by improving data generation and minimizing the need for extensive online interactions, thereby driving significant advancements in the effectiveness and applicability of offline RL solutions [53, 14]. By addressing these challenges and exploring new methodologies, offline RL can continue to expand its impact across various sectors, ensuring its methodologies remain relevant and effective in tackling the complexities of modern dynamic environments.

# 4 Decision Transformer (DT)

## 4.1 Introduction to Decision Transformer

The Decision Transformer (DT) redefines Reinforcement Learning (RL) by conceptualizing RL tasks as sequence modeling problems, employing transformer architectures to streamline the RL process by modeling distributions over trajectories [41]. By treating states, actions, and rewards as unstructured sequences, DT enhances policy learning efficiency and scalability across diverse tasks. This hierarchical framework allows for breaking down complex tasks into subtasks, facilitating incremental problem-solving and reuse of solutions, akin to human decision-making [36].

DT incorporates a dual-policy mechanism, inspired by Recovery RL, which optimizes task performance while maintaining constraint adherence through a recovery policy [38]. This enhances robustness, particularly in safety-critical environments. DT's significance is highlighted by its integration of conditional policy approaches with transformer architecture, achieving competitive performance in offline RL tasks. However, limitations in stitching capabilities restrict optimal policy learning from sub-optimal trajectories, especially when datasets lack diversity. Innovations like the Q-learning Decision Transformer (QDT) merge dynamic programming strengths with the DT framework, improving real-time decision-making [13]. Language model-initialized Prompt Decision Transformers (LPDT) further enhance few-shot learning and task differentiation, reinforcing DT's potential across various contexts [15].

## 4.2 Transformer Architectures in Sequence Modeling

Transformer architectures have been adapted for sequence modeling in RL, leveraging attention mechanisms to manage sequential data. This enables the modeling of complex dependencies among states, actions, and rewards in decision-making. Utilizing the distributional Bellman operator, these architectures approximate intricate reward distributions within Banach spaces, establishing a robust theoretical foundation [66]. The self-attention mechanism captures long-range dependencies, crucial for tasks requiring temporal dynamics understanding. Models like the Decision Transformer enhance performance by using trajectory portions as prompts. Despite high computational costs due to quadratic complexity, advancements like the Decision Mamba-Hybrid model merge transformer strengths with efficient memory processing, improving computational efficiency [15, 67].

By framing RL problems as sequence modeling tasks, transformers integrate information across trajectories, leading to informed policy decisions, enhancing scalability and generalization across varied tasks. Their adaptability positions them as valuable assets in RL applications with high-dimensional state spaces and complex reward structures. Transformers model distributions over trajectories, facilitating effective policy learning from limited datasets and enabling in-context reinforcement learning, crucial for dynamic environments. Innovations like the Q-learning Decision Transformer improve performance with sub-optimal trajectories, advancing modern RL methodologies by addressing generalization, stability, and computational efficiency challenges [67, 68].

### 4.3 Benefits of Decision Transformer in RL

The Decision Transformer (DT) offers numerous advantages by reconceptualizing RL tasks as sequence modeling problems, utilizing transformer architectures to manage long-term dependencies and complex decision-making tasks [41]. DT unifies RL components into a single model, enhancing scalability and effectiveness in long-horizon tasks, promoting efficient planning over high-dimensional state spaces [41]. DT's subtask-driven approach, exemplified by the RED RL framework, enhances efficiency and robustness, improving generalization capabilities [26]. Incorporating knowledge-based systems reduces data requirements and fosters explainability, enhancing trust in AI systems [22].

DT aligns with advanced RL methodologies emphasizing exploration-exploitation balance. Recovery RL balances task performance and safety, allowing safer exploration and more efficient learning [38]. DT's adaptability is evident through integration with methods like SFOLS, ensuring optimal policy identification for linearly-expressible tasks, outperforming traditional methods [29]. DT can leverage Monte Carlo Tree Search (MCTS) for improved decision-making [51].

Despite challenges in stitching sub-optimal trajectories, innovations like QDT integrate dynamic programming to enhance policy learning. LPDT improves few-shot learning and task differentiation, positioning DT and its derivatives as powerful tools for modern RL applications, addressing data efficiency and performance challenges [13, 15].

### 4.4 Limitations and Challenges

The Decision Transformer (DT) framework encounters several limitations impacting performance and applicability. Computational overhead of transformer architectures can lead to increased training times and resource consumption, limiting real-time application feasibility [30]. Frequent calls to guiding policies, as in Monte Carlo Tree Search (MCTS), extend training duration [51]. DT's reliance on complete trajectories for returns computation poses challenges in fragmented data environments [45]. Quality and completeness of offline datasets are crucial for DT's effectiveness; methods like QDT may struggle with suboptimal trajectories [13]. Dependency on encoded human knowledge in Knowledge-Based Reinforcement Learning (KB-RL) may not reflect optimal strategies across scenarios [22].

This figure illustrates the limitations and challenges faced by the Decision Transformer framework, categorizing these challenges into computational challenges, data dependence, and adaptation and evaluation issues, while highlighting key references that delve into these aspects Figure 4. Reward hacking, where models exploit metrics without improving human-aligned performance, poses challenges, highlighting the need for robust evaluation metrics [69]. Effective out-of-distribution (OOD) detection mechanisms are essential for policy stability in dynamic environments [70]. DT struggles with adaptation to dissimilar environments, emphasizing the need for adaptive models capable of generalizing across diverse environments [71]. Bootstrapped Transformer approach enhances performance but increases training time, limiting practicality in time-sensitive applications [14]. Addressing these challenges is vital for enhancing DT's effectiveness in RL tasks, particularly in economics with high-dimensional data and distributional shifts [72].

### 4.5 Recent Research and Applications

Recent advancements in RL highlight the Decision Transformer (DT) model's versatility across various applications. The In-Context Decision Transformer (IDT) has demonstrated superior adaptability and performance in managing complex sequential decision-making tasks across diverse task horizons, outperforming baseline methods like Agentic Transformer (AT) and Algorithm Distillation (AD) [73].

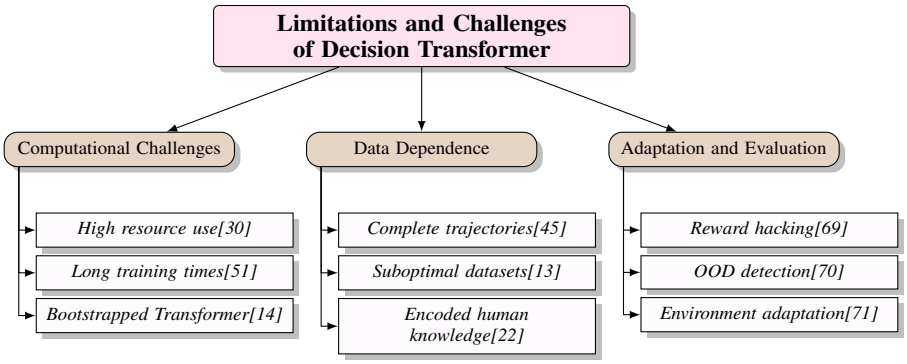


Figure 4: This figure illustrates the limitations and challenges faced by the Decision Transformer framework. It categorizes these challenges into computational challenges, data dependence, and adaptation and evaluation issues, highlighting key references that delve into these aspects.

The Decision Mamba (DM-H) method further exemplifies DT models' potential, showcasing their capacity to generalize across different tasks [67]. DT models have been applied to domains including Atari games and robotics tasks, demonstrating practical utility and robust model development for complex control problems [74].

DRL, including DT models, extends to financial domains like stock trading and portfolio management, optimizing strategies and enhancing online services [72]. Pretrained language models improve decision-making in RL tasks, with models like LPDT enhancing performance in environments like MuJoCo control tasks and Meta World ML1 [15]. These research efforts illustrate DT's growing impact in advancing RL capabilities, addressing challenges like inconsistent evaluation metrics and data-hungry models, ultimately leading to more effective algorithm development across diverse applications [15, 75].

# 5 Deep Reinforcement Learning

## 5.1 Challenges in Deep Reinforcement Learning

Deep Reinforcement Learning (Deep RL) presents significant challenges, particularly in integrating deep learning with reinforcement learning frameworks. Continuous control tasks demand extensive training data and computational resources, complicating real-time applications. The curse of dimensionality exacerbates this by expanding the state-action space, leading to value and policy churn that can bias learning dynamics and reduce performance. Additionally, the non-stationary nature of RL training complicates these issues, requiring robust strategies to navigate interactions within high-dimensional reward spaces [66, 76, 77, 78].

The exploration-exploitation dilemma remains central in model-free RL, necessitating a balance between exploring novel actions and exploiting known rewarding actions. This balance is vital for effective learning in high-dimensional spaces. The opacity of RL policies, especially those using deep neural networks, hinders interpretability and transparency, complicating deployment in critical applications. Enhancing explainability through intrinsic interpretability and post-hoc explanations is increasingly emphasized to clarify complex models, fostering trust and improving RL agents' effectiveness in real-world scenarios [79, 52, 78].

As illustrated in Figure 5, the primary challenges in Deep Reinforcement Learning can be categorized into three main areas: integration challenges, the exploration-exploitation dilemma, and optimization and reward design issues. Each category highlights key aspects and references relevant studies addressing these challenges. Reward function quality significantly influences learning dynamics, potentially leading to suboptimal policy development in complex environments. This highlights the need for careful reward design and evaluation, as variations can dramatically affect RL algorithm performance across applications like dynamic treatment regimes and recommendation systems [55, 80, 78]. Robust reward functions are crucial, especially in complex settings where suboptimal structures can hinder learning. Additionally, Reinforcement Policy Optimization (RPO) sensitivity

to perturbation parameters requires careful tuning for optimal performance, adding complexity to training.

Selecting an appropriate learning rate during deep RL model training is another challenge, as unsuitable rates can impair learning effectiveness. On-policy RL optimization is further complicated by initial action distribution and design choices influencing agent performance, with over 50 critical design decisions affecting RL outcomes across continuous control tasks [81, 77, 82, 83, 84].

Introducing entropy in optimization methods can enhance exploration but may initially slow training due to increased randomness in action selection [28]. However, advancements like the Trajectory Transformer, which generates long-horizon trajectories, show potential for overcoming these hurdles in offline RL tasks. Methods like SFOLS outperform existing algorithms in classic multi-objective reinforcement learning (MORL) and successor feature (SF) domains, offering promising solutions for managing complex RL tasks [29].

Addressing Deep RL challenges is vital for progress, enhancing the ability to leverage large datasets and tackle complex decision-making scenarios. This includes developing benchmarks for offline RL, mitigating value and policy churn, and incorporating innovative frameworks like Language Guided Exploration to improve learning efficiency and adaptability in dynamic environments. By overcoming these obstacles, researchers can enhance Deep RL methods' robustness and applicability in various practical contexts, including economics and logistics [72, 52, 6, 85, 78]. Focusing on computational efficiency, exploration-exploitation balance, reward function design, and learning rate optimization will continue to push Deep RL boundaries.

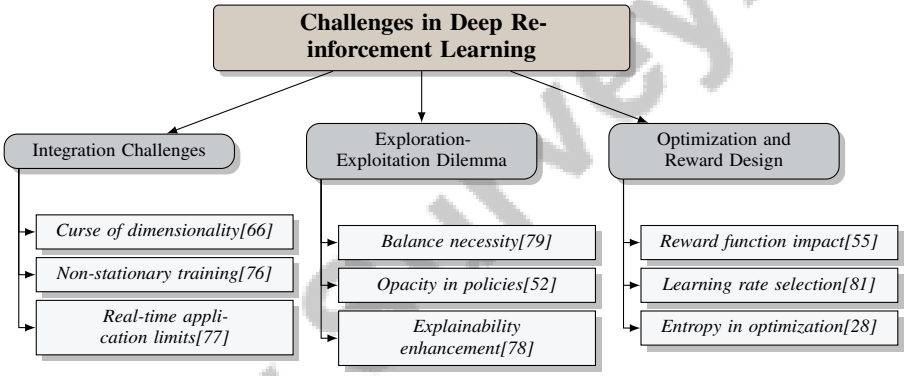


Figure 5: This figure illustrates the primary challenges in Deep Reinforcement Learning, categorizing them into integration challenges, the exploration-exploitation dilemma, and optimization and reward design issues. Each category highlights key aspects and references relevant studies addressing these challenges.

### 5.2 Innovative Methods to Enhance Sample Efficiency

Enhancing sample efficiency in Deep Reinforcement Learning (Deep RL) is critical due to extensive data requirements. Several innovative methods address this challenge, focusing on improving models' efficiency in learning from available data. Smooth Regularized Reinforcement Learning (SR2L) learns a smooth policy maintaining consistent behavior across similar states, enhancing sample efficiency and robustness by reducing policy variability [86].

The hybrid actor-critic framework integrates broad learning systems (BLS) for the critic network with deep neural networks (DNN) for the actor network, enhancing learning efficiency without compromising accuracy by leveraging broad learning and deep learning techniques to optimize policy performance [87].

Optimizing the reward mechanism is crucial for improving sample efficiency. Directly maximizing the average reward criterion, as opposed to the traditional discounted reward approach, provides a more robust framework for learning, potentially leading to more efficient policy optimization by emphasizing long-term average performance over immediate rewards [88].

Dynamic adjustment of learning parameters offers another avenue for enhancing sample efficiency. The dynamic Learning Rate for deep Reinforcement Learning (LRRL) employs a multi-armed bandit

algorithm to select learning rates on-the-fly, adapting to the agent's performance during training. This approach allows for more responsive and efficient learning by dynamically tuning the learning rate to suit the current training context [89].

The Sample Mean Representation (SMR) method provides a straightforward yet effective way to boost sample efficiency. By incorporating SMR into existing algorithms without complex modifications, researchers can achieve significant improvements in data utilization and learning outcomes [59].

Additionally, the Churn Approximated ReductIoN (CHAIN) method minimizes undesirable changes in policy and value networks' outputs for states not included in the current training batch, enhancing stability and sample efficiency, contributing to more effective learning processes [78].

Future research directions include integrating these innovative methods with other RL techniques to enhance adaptability and performance across diverse problem scenarios. By combining these approaches, researchers aim to develop more robust and efficient RL models capable of tackling complex tasks with greater data efficiency [90].

### 5.3 Integration of Deep Learning with Control Theory

Integrating deep learning with control theory in Reinforcement Learning (RL) marks a significant advancement, enhancing capabilities for managing complex, high-dimensional state spaces. This synergy leverages deep learning's ability to process vast data and learn intricate patterns alongside control theory's robust mathematical frameworks, essential for designing stable and efficient control systems [91].

Deep learning techniques, particularly deep neural networks, extend RL's applicability to environments characterized by high-dimensional state spaces, enabling feature extraction and approximation of complex value functions and policies, crucial for effective decision-making in dynamic environments. Deep learning enhances RL algorithms' capacity to generalize across diverse tasks, providing a flexible framework for tackling a wide range of control problems [91].

Integrating control theory into RL frameworks, such as the proposed SR²L approach, offers a systematic methodology for enhancing learned policies' stability and robustness. This integration leverages smoothness-inducing regularization to constrain the search space and improve sample efficiency while making policies more resilient to measurement errors in continuous state environments. It allows for natural extensions into distributionally robust settings, leading to more effective performance in various continuous control tasks [86, 77]. Control theory provides tools for analyzing and designing systems that maintain desired performance levels despite uncertainties and perturbations. By integrating these tools with deep learning, researchers can develop RL algorithms that not only learn optimal policies but also adhere to stability and robustness criteria, critical for applications in safety-critical domains.

Innovative methodologies like LRRL exemplify integrating control principles with deep learning techniques. LRRL utilizes a meta-learning approach to dynamically select the learning rate during training, employing a multi-armed bandit algorithm based on RL policy's cumulative returns. This adaptive strategy enhances learning by ensuring the learning rate is optimized for the current training context, improving convergence and overall performance [89].

Regularization methods like CHAIN contribute to integrating control theory by minimizing value and policy churn in deep RL. CHAIN achieves this by reducing changes in target values for a separate batch of data during training, enhancing learned policies' stability and reliability [78].

Integrating deep learning with control theory in RL is an evolving research domain characterized by ongoing efforts to enhance methodologies addressing challenges like value and policy churn, the need for specialized benchmarks in offline RL, and optimizing hybrid architectures for continuous control tasks. Researchers are refining these approaches to improve learning performance and computational efficiency, as evidenced by advancements like the CHAIN method, tailored offline RL benchmarks, and innovative architectures combining broad learning systems with deep neural networks [87, 52, 78]. By merging both fields' strengths, researchers aim to develop more robust, efficient, and scalable RL algorithms capable of addressing modern control systems' complex challenges.

### 5.4 Hierarchical and Multi-Objective Approaches

Hierarchical and multi-objective approaches in Deep Reinforcement Learning (Deep RL) are powerful methodologies for addressing high-dimensional and dynamic environments' complexities. Hierarchical reinforcement learning (HRL) simplifies complex tasks by breaking them into manageable subtasks, enhancing learning efficiency. This approach allows agents to focus on specific problem components, simplifying the credit-assignment challenge and improving data efficiency. Recent advancements, like the HIRO algorithm, optimize HRL by enabling off-policy experience use for training both higher- and lower-level policies, significantly reducing interactions required for effective learning. This method streamlines learning and facilitates meaningful sub-goals' automatic generation, leading to more sophisticated and adaptable behaviors in real-world applications, like robotic control [92, 93, 79]. This decomposition aligns with human-like decision-making processes, where strategic planning is often broken into sequential steps, enabling effective exploration and exploitation of the environment.

A notable advancement in this area is SR2L, ensuring policy stability by minimizing drastic changes in response to small state perturbations. This approach enhances RL models' robustness and performance by maintaining consistent behavior across similar states, critical for reliable learning outcomes [86].

In multi-objective RL, methods like Multi-objective Q-Learning with Global Statistics (MOSS) enhance traditional Q-learning frameworks by leveraging accumulated global statistics to inform action selection, improving decision-making in environments with multiple competing objectives [94]. This approach allows for a nuanced balance between different objectives, enabling agents to optimize performance across various task dimensions.

Hierarchical and multi-objective strategies in Deep RL are further supported by adaptive techniques like LRRL. LRRL employs a meta-learning approach to dynamically adjust learning rates based on agent performance, enhancing adaptability and efficiency in learning processes. However, this method may still require fine-tuning of initial learning rates and may not achieve optimal performance across all environments or tasks, emphasizing careful parameter selection [89].

Ongoing investigations into hierarchical and multi-objective strategies within Deep RL significantly enhance the field by providing sophisticated solutions tailored for complex decision-making scenarios, particularly in dynamic environments characterized by high-dimensional data, non-linearity, and uncertainty. These advancements improve DRL applications' accuracy and performance in areas like economics and inventory management while addressing challenges like value and policy churn, facilitating more effective learning and decision-making processes [95, 72, 6, 78]. By effectively decomposing tasks and balancing multiple objectives, these methodologies enhance Deep RL's scalability and applicability across diverse domains, paving the way for more sophisticated and capable RL systems.

### 5.5 Notable Deep RL Algorithms and Applications

Deep Reinforcement Learning (Deep RL) has significantly advanced through innovative algorithms enhancing performance in training stability, sample efficiency, and applicability across diverse domains. A notable contribution is the decorrelation method, improving Deep Q-Network (DQN) and Quantile Regression DQN (QR-DQN) algorithms' performance, achieving superior or comparable results across most games [19]. This method effectively addresses correlated data challenges, impeding learning efficiency in RL environments.

Proximal Policy Optimization (PPO) has emerged as a robust algorithm, consistently outperforming traditional methods like Deep Q-Learning (DQN) in various tasks. The Accelerated Proximal Policy Optimization (APPO) approach enhances learning speed and final performance, particularly in environments with delayed rewards, demonstrating efficacy compared to standard PPO and offline-only methods [1]. This advancement underscores adaptive algorithms' importance in handling complex control tasks.

The Trajectory Transformer, evaluated using the D4RL offline benchmark suite, demonstrates competitive performance against state-of-the-art offline RL methods based on cumulative rewards [30]. This model exemplifies transformer-based architectures' potential in optimizing decision-making processes across various RL tasks.

The Neurally Approximated Lyapunov Optimal Control method, tested on linear and nonlinear control tasks like Double Integrator and Cartpole balancing, shows favorable comparisons against Soft Actor Critic (SAC) and Proximal Policy Optimization (PPO) [31]. This approach highlights integrating control theory with deep learning techniques to enhance RL performance in dynamic environments.

The Discrete-to-Deep Supervised Policy Learning (D2D-SPL) method, applied to RL environments like the Cartpole problem and an aircraft maneuvering simulator, demonstrates effectiveness against various baseline methods, including DQN, DDQN, and A3C [20]. This method showcases transforming discrete state spaces into continuous ones, improving learning efficiency and policy performance.

Deep Reinforcement Learning advancements are exemplified by innovative algorithms and applications showcasing its potential to enhance AI systems across diverse domains. The introduction of offline RL benchmarks facilitates utilizing large, previously collected datasets, akin to supervised learning progress. Integrating Large Language Models in the Language Guided Exploration framework enables RL agents to navigate complex decision-making environments more effectively. Additionally, Deep RL applications in economics demonstrate its capability to tackle high-dimensional problems characterized by noisy and nonlinear data patterns, while new methods like CHAIN address challenges in RL training dynamics. Collectively, these developments highlight Deep RL's transformative impact, from optimizing control systems to mastering competitive gameplay [72, 6, 52, 78]. Continuous exploration and refinement of these methodologies promise to expand Deep RL's impact, paving the way for more sophisticated and capable RL systems in the future.

# 6 Policy Optimization

## 6.1 Conceptual Framework and Techniques

The framework of policy optimization in Reinforcement Learning (RL) is central to enhancing decision-making across tasks, especially in dynamic environments. It encompasses techniques that refine RL policies, improving adaptability and performance by using semantically meaningful skills from offline data without supervision and minimizing evaluation error penalties in Q-function learning [76, 96]. Incorporating human knowledge is crucial for enhancing explainability and performance, as highlighted by the need for clear taxonomy in Explainable Reinforcement Learning (XRL) [79], fostering trust and transparency.

The development of robust controllers that learn directly from data is another critical aspect, with methods showing improved performance under varying conditions compared to traditional approaches [97]. Adaptive policies, capable of managing uncertainty and providing optimal solutions in Markov Decision Processes (MDPs), represent a significant strength in contemporary research [98]. These policies are essential for navigating uncertainties, enabling informed decision-making amidst incomplete information.

Incorporating Knowledge Representation and Reasoning (KRR) methods into RL frameworks enhances learning efficiency, policy generalization, and safety [9]. By leveraging domain knowledge, RL agents achieve more efficient learning and improved generalization across diverse tasks. The Drift Policy Optimization (DPO) algorithm exemplifies a closed-form approach that employs a drift function for policy updates, ensuring convergence to optimal policies [99].

Viewing RL problems through the lens of duality, particularly in dynamic programming and policy optimization, provides a novel perspective that can improve the effectiveness of RL algorithms [33]. In constrained RL settings, minimizing improvement penalties by constraining policy updates is crucial to prevent actions with low sampling frequency from disproportionately affecting policy performance [96].

The RED RL method minimizes Temporal Difference (TD) error, enabling simultaneous optimization of multiple subtasks, which highlights the potential of subtask-driven approaches in enhancing policy optimization [26]. However, manual threshold parameter selection for -lexicographic constraints requires careful estimation [36]. The framework for policy optimization in RL emphasizes robustness, adaptability, and efficiency, with advancements like Robust Policy Optimization (RPO) enhancing policy exploration through high-entropy actions and Smooth Regularized Reinforcement Learning ($SR^2L$) enforcing smoothness in policy behavior for improved sample efficiency and robustness against measurement errors [86, 84, 77, 96].

### 6.2 On-Policy vs. Off-Policy Methods

On-policy and off-policy methods are fundamental approaches in policy optimization within RL, each employing distinct strategies for learning and updating policies. On-policy methods optimize decision-making based on real-time feedback from actions taken by the current policy, enhancing stability and convergence, though often suffering from limited sample efficiency [62, 76, 77]. In contrast, off-policy methods enhance sample efficiency by leveraging data from previous or different policies, allowing for broader exploration without requiring constant interaction with the environment. This approach is advantageous in scenarios where data collection is costly or risky, facilitating policy optimization through pre-existing datasets [61, 58, 100, 60].

The handling of Q-value estimation inaccuracies further distinguishes these approaches. On-policy methods are more sensitive to inaccuracies due to their reliance on immediate feedback, which can degrade performance if Q-values are misestimated [96]. Off-policy methods can mitigate some inaccuracies by incorporating diverse data sources, though they still encounter challenges related to stability and reliability of policy updates.

Leveraging duality principles in RL algorithms presents a promising avenue for addressing limitations in both on-policy and off-policy methods, potentially improving convergence properties and interpretability [33]. The choice between on-policy and off-policy methods depends on task requirements, including sample efficiency, data availability, and the balance between exploration and exploitation. Language Guided Exploration (LGE) and Deep Reinforcement Learning (DRL) methodologies offer distinct advantages in improving RL agent performance, with LGE enhancing learning efficiency in complex environments and DRL handling high-dimensional data [72, 9, 6].

### 6.3 Robustness and Stability in Policy Optimization

Robustness and stability are crucial in policy optimization within RL, significantly affecting the reliability and performance of learned policies in dynamic and uncertain environments. Recent advancements have introduced methods that guarantee convergence in infinite-state settings, marking a substantial step forward in maintaining stability across diverse state spaces [101]. Safe policy improvement techniques, like the Risk-Bounded Improvement (RBI) framework, provide safety advantages during policy updates, particularly in scenarios with high variance optimal actions [96]. Constrained policy improvement methods focus on minimizing the impact of inaccurate Q-value estimations, preventing suboptimal policy performance [96].

Integrating duality principles into RL algorithms offers additional pathways for enhancing robustness and stability, improving convergence properties and interpretability [33]. A comprehensive strategy is essential, integrating cutting-edge optimization methods, safe policy enhancement frameworks, and novel theoretical insights. For instance, Robust Policy Optimization (RPO) enhances policy entropy and maintains exploratory behavior, while frameworks like Smooth Regularized Reinforcement Learning (SR²L) improve sample efficiency and robustness against measurement errors. Incorporating uncertainty estimation techniques in algorithms like Proximal Policy Optimization (PPO) enhances agents' safety and reliability by enabling the identification of out-of-distribution states. Addressing corruption robustness in offline RL through uncertainty-weighted approaches ensures policies remain effective amidst adversarial data corruption [86, 81, 84, 102].

### 6.4 Advanced Techniques and Innovations

Advanced techniques and innovations in policy optimization have significantly advanced RL, particularly in enhancing the efficiency and robustness of policy learning frameworks. A notable innovation is the decoupled policy structure within the Soft Actor-Critic framework, which separates the mean and deviation of the Gaussian policy, reducing computational demands and improving efficiency in high-dimensional spaces [103]. The convergence rate of $O(1/\sqrt{T})$ for the Natural Policy Gradient (NPG) algorithm in infinite-state average-reward Markov Decision Processes (MDPs) represents a significant advancement in ensuring stable and reliable policy optimization [101].

In Figure 6, we illustrate the advanced techniques and innovations in reinforcement learning (RL), categorized into policy optimization methods, constrained and robust RL approaches, and algorithm discovery. Each category highlights key methods contributing to the efficiency and robustness of RL frameworks. Simultaneous Perturbation Stochastic Approximation (SPSA) algorithms enhance

computational efficiency by providing unbiased gradient and Hessian estimates using only two evaluations, crucial for optimizing policies in environments with limited computational resources [104]. The integration of reward-free approaches in constrained RL settings streamlines the optimization process, achieving efficient solutions while reducing algorithmic complexity [105].

Incorporating multi-objective optimization techniques, such as those used in robust model-free RL methods, allows for simultaneous optimization of performance and robustness indicators. Multi-objective Bayesian optimization exemplifies this approach, enhancing the overall effectiveness of RL algorithms [97]. The Discovered Policy Optimization (DPO) algorithm stands out for its theoretical soundness, ease of implementation, and superior performance compared to existing methods like Proximal Policy Optimization (PPO), particularly in unseen environments [99].

Finally, the Risk-Bounded Improvement (RBI) framework ensures safe policy updates by defining behavior policies and applying reroute constraints, analyzing improvement penalties based on estimated Q-functions, and calculating optimal policies using linear programming [96]. These advancements collectively underscore ongoing innovations in policy optimization, paving the way for more robust and efficient RL systems.

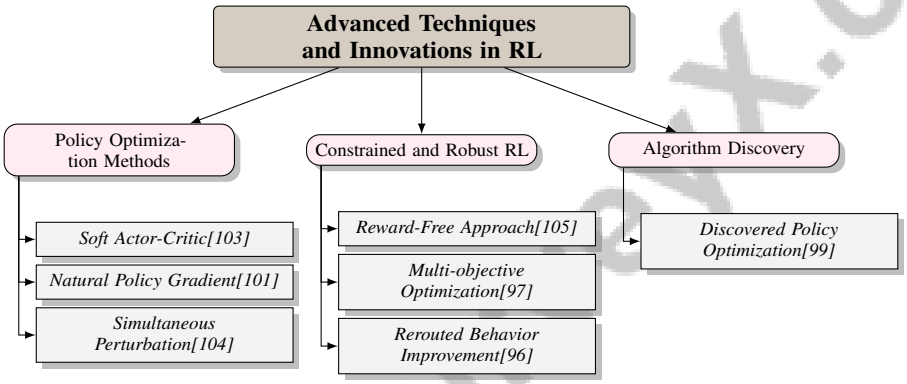


Figure 6: This figure illustrates the advanced techniques and innovations in reinforcement learning (RL), categorized into policy optimization methods, constrained and robust RL approaches, and algorithm discovery. Each category highlights key methods contributing to the efficiency and robustness of RL frameworks.

# 7 Model-Free Methods

## 7.1 Introduction to Model-Free Methods

Model-free methods in Reinforcement Learning (RL) are pivotal for learning optimal policies or value functions directly from experience, eliminating the need for explicit environmental models [16]. This is advantageous in complex environments where modeling dynamics is impractical. Algorithms like Q-learning and Policy Gradient methods optimize expected cumulative rewards by iteratively updating value estimates or policy parameters based on observed rewards and transitions.

These methods excel in handling high-dimensional and continuous action spaces, making them suitable for diverse applications, including robotics and autonomous systems [106]. Deep neural networks often approximate value functions or policies, allowing intricate mappings from states to actions in dynamic environments [19].

Balancing exploration and exploitation is a core challenge in model-free RL, especially in environments with sparse rewards and large decision spaces. Innovative strategies like intrinsically-motivated RL and frameworks such as Language Guided Exploration (LGE) enhance this balance by using intrinsic rewards and pre-trained language models to guide decision-making [6, 105, 2]. Techniques such as epsilon-greedy exploration and intrinsic motivation improve the efficiency and robustness of these algorithms.

Despite their successes, model-free methods face challenges with sample inefficiency, often requiring extensive trials to identify effective policies. Recent advancements focus on enhancing sample effi-

ciency through bootstrapped representations and intrinsic motivation strategies, including exploration bonuses derived from novelty and curiosity. Integrating techniques like Rényi state entropy maximization provides sustainable exploration incentives, enabling agents to learn efficiently from limited data. Learning exploration bonuses from demonstrations facilitates complex exploration behaviors, enhancing performance in tasks requiring efficient learning without exhaustive exploration [2, 107]. These developments highlight the ongoing evolution of model-free methods and their significance in advancing RL capabilities across various domains.

### 7.2 Handling Non-Stationarity in Model-Free RL

Managing non-stationarity in model-free RL is crucial due to the dynamic nature of many real-world environments, where transition dynamics or reward functions may change over time. Such non-stationarity can adversely affect the performance and stability of RL algorithms, as policies learned in one context may become suboptimal when the environment changes [16].

Adaptive learning algorithms that adjust to environmental changes are essential for managing non-stationarity. These algorithms often include mechanisms for detecting shifts and updating policies accordingly. Bootstrapped representations, for instance, utilize ensemble learning to maintain multiple hypotheses about the environment, allowing agents to adaptively switch among them as conditions evolve [18].

Leveraging intrinsic motivation to promote exploration in dynamic environments is another strategy. By incorporating intrinsic rewards independent of the external environment, agents sustain exploration levels that facilitate adaptation to changes, enhancing the robustness of model-free RL algorithms in non-stationary settings [2].

Meta-learning and continual learning techniques also improve adaptability. Meta-learning frameworks enable agents to learn how to learn, quickly adapting to new tasks by leveraging prior knowledge. Continual learning focuses on retaining knowledge from past experiences while adapting to new ones, mitigating the effects of catastrophic forgetting when environments change [21].

A multifaceted approach is essential for effectively tackling non-stationarity in model-free RL, integrating adaptive learning strategies, exploration techniques that balance pre-trained offline policies with online policies for enhanced data efficiency, and knowledge retention mechanisms. This comprehensive strategy improves cumulative rewards and computational efficiency while aligning with established theoretical performance benchmarks [62, 108]. Researchers aim to develop more resilient RL algorithms capable of maintaining performance amidst dynamic and unpredictable environmental changes.

### 7.3 Key Algorithms in Model-Free RL

| Method Name | Learning Approach | Action Space Suitability | Stability Enhancements |
|---|---|---|---|
| DQN-decor[19] | Policies Directly | Discrete | Experience Replay |
| RISE[2] | Intrinsic Rewards | Discrete And Continuous | Experience Replay |
| APPO[1] | Hybrid Policy Architecture | Delayed Rewards Environments | Reward Shaping Mechanism |

Table 1: Overview of selected model-free reinforcement learning algorithms, highlighting their learning approaches, action space suitability, and stability enhancements. This table provides insights into the distinct strategies employed by DQN-decor, RISE, and APPO to improve learning performance in various environments.

Model-free RL algorithms are key to learning optimal policies or value functions directly from environmental interactions without requiring explicit models of dynamics. Q-learning, a prominent model-free algorithm, iteratively updates the value of action-state pairs based on the Bellman equation, facilitating optimal policy learning through temporal difference learning [16]. This method excels in discrete action spaces, where Q-value function approximation is efficient.

Deep Q-Networks (DQN) extend Q-learning to manage high-dimensional state spaces by employing deep neural networks for Q-value function approximation. DQN incorporates experience replay and target networks to stabilize learning, addressing correlation issues in sequential data and non-stationarity in policy updates [19]. These innovations have significantly enhanced model-free RL performance in complex environments, such as video games and robotic control tasks.

Policy Gradient methods, including the REINFORCE algorithm, optimize policies directly by estimating the gradient of the expected reward concerning policy parameters, making them particularly suitable for continuous action spaces where value-based methods like Q-learning may struggle [2]. The Proximal Policy Optimization (PPO) algorithm refines policy gradient methods by introducing a surrogate objective function that ensures stable and efficient policy updates, thus becoming a popular choice for various RL tasks [1].

Actor-Critic methods, such as Advantage Actor-Critic (A2C) and its asynchronous variant A3C, merge value-based and policy-based approaches by maintaining separate networks for policy and value estimation. These methods utilize the critic to provide a baseline for reducing variance in policy gradient estimates, enhancing learning stability and efficiency [2].

Recent advancements in model-free RL emphasize improving sample efficiency and exploration strategies. Techniques like bootstrapped representations and intrinsic motivation have been developed to encourage exploration and enhance data utilization, addressing challenges associated with sparse rewards and high-dimensional state spaces. These innovations, including reward-free RL and language-guided exploration, significantly advance model-free RL, enabling agents to function effectively in complex decision-making environments while addressing safety and multi-objective constraints [6, 105]. Table 1 presents a comparative analysis of key model-free reinforcement learning algorithms, elucidating their respective learning approaches, action space applicability, and mechanisms for enhancing stability.

### 7.4 Comparison with Model-Based Methods

Model-free and model-based RL methods present distinct approaches to sequential decision-making tasks, each with unique advantages and limitations. Model-free methods, such as Q-learning and Policy Gradient techniques, focus on learning optimal policies directly from environmental interactions without explicit models of dynamics. These methods excel in high-dimensional and complex environments where modeling dynamics is infeasible or computationally burdensome [106].

Conversely, model-based methods involve creating predictive models that capture environmental dynamics, which are then used to simulate potential future states and associated rewards, enhancing decision-making processes. These methods optimize agent behavior through decision-time and background planning approaches. By employing learned models, agents can generate controllable policies, adapt to environmental constraints, and achieve near-optimal performance with minimal real-time interaction, as evidenced in tasks such as battery management and robotics. The choice of underlying models—ranging from neural networks to Gaussian processes—significantly impacts the effectiveness of these reinforcement learning algorithms, emphasizing the importance of model selection tailored to specific applications [109, 110, 111, 112]. This approach allows for more sample-efficient learning by enabling planning and policy evaluation using simulated experiences rather than relying solely on real-world interactions.

A key difference between the two approaches lies in their exploration-exploitation balance; model-free methods typically utilize intrinsic motivations for exploration, while model-based methods leverage learned models to plan and evaluate multiple potential actions before execution [107, 82, 6, 2]. This capability can lead to more informed decision-making and potentially faster convergence to optimal policies.

Despite their advantages, model-based methods face challenges related to model accuracy and computational complexity. Learning an accurate model can be difficult, particularly in environments with stochastic dynamics or partial observability. The computational overhead associated with planning and model updates in model-based reinforcement learning can be substantial, posing challenges for real-time or resource-constrained implementations. Recent advancements, such as Model-Based Offline Planning (MBOP), aim to mitigate these issues by generating controllable policies directly from offline data, reducing reliance on real-time interactions and enhancing performance with minimal system engagement. However, the trade-off between computational efficiency and model update complexity remains a critical consideration for effectively applying these techniques in dynamic settings [109, 110, 111].

Recent developments in RL have focused on integrating the strengths of both model-free and model-based approaches to create hybrid methods. These hybrid techniques seek to enhance learning efficiency and robustness by leveraging model-free strategies that prioritize performance in dynamic

environments while incorporating model-based elements to improve sample efficiency and reduce unnecessary exploration. For instance, studies have demonstrated that hybrid inverse reinforcement learning can effectively combine online and expert data, minimizing computational waste and maximizing learner focus on relevant states. Additionally, robust model-free RL is being explored through multi-objective optimization, addressing performance and robustness challenges in real-world applications where training conditions differ from test scenarios. Collectively, these advancements signify a promising direction in RL research, aimed at producing more resilient and efficient learning algorithms [97, 113]. These hybrid approaches leverage the direct policy learning capabilities of model-free methods while incorporating model-based planning to enhance sample efficiency and decision-making accuracy.

### 7.5 Applications and Practical Considerations

| Method Name | Application Domains | Challenges and Solutions | Methodological Innovations |
|---|---|---|---|
| HDL-CRM[3] | Customer Relationship Management | Hidden State Inference | Joint Training |
| RISE[2] | Autonomous Systems | Sample Inefficiency | Intrinsic Motivation |
| RBI[96] | Atari Learning Environment | Sample Inefficiency, Policy Robustness | Reroute Constraint |

Table 2: Overview of selected model-free reinforcement learning methods, highlighting their application domains, challenges encountered, and methodological innovations. The table provides insights into how these methods address issues such as sample inefficiency and policy robustness, contributing to advancements in diverse fields such as customer relationship management, autonomous systems, and gaming.

Model-free RL methods have broad applications across various domains, capitalizing on their ability to learn directly from environmental interactions without requiring explicit models of dynamics. A prominent application is in robotics, where model-free methods develop control policies for complex tasks such as manipulation and locomotion [106]. Their capability to manage high-dimensional state and action spaces makes model-free RL particularly suitable for robotics, where precise control and adaptability are essential.

In autonomous systems, model-free RL enhances decision-making in uncertain and dynamic environments. For instance, these methods optimize strategies in autonomous driving, where safety and efficiency are critical [12]. The flexibility of model-free approaches allows for robust policy development that adapts to varying traffic conditions and unforeseen obstacles, making them invaluable for real-time decision-making.

Model-free RL is also applied in finance to optimize trading strategies and portfolio management. By learning directly from market data, model-free algorithms identify patterns and make informed decisions to maximize returns while managing risk [72]. Their ability to operate in high-frequency trading environments, where rapid decision-making is crucial, underscores the utility of model-free methods in finance.

In healthcare, model-free RL optimizes treatment strategies and personalizes patient care. By learning from patient data, these methods recommend optimal treatment plans that enhance outcomes while minimizing costs [3]. The adaptability of model-free approaches is particularly beneficial in healthcare, where individual patient responses can vary significantly, necessitating personalized strategies.

Despite their versatility, model-free RL methods face practical challenges that must be addressed for successful deployment. A key issue is sample inefficiency, as these methods often require extensive data to learn effective policies, which can be costly and time-consuming to obtain [2]. Techniques like intrinsic motivation and bootstrapped representations have been developed to enhance exploration and improve data utilization [18].

Another consideration is the stability and robustness of learned policies, especially in safety-critical applications. Ensuring policies remain reliable under varying conditions and noise is essential for practical implementation. This necessitates robust optimization techniques and safety constraints to guide policy learning and execution [96].

Model-free RL methods hold significant potential for advancing decision-making capabilities across diverse applications. By addressing critical challenges such as sample efficiency and policy robustness, these advanced methods enhance effectiveness in solving complex real-world problems. Techniques

like uncertainty estimation in Proximal Policy Optimization (PPO) improve safety and reliability by enabling agents to quantify uncertainty and detect out-of-distribution states. The EXTRACT method utilizes pre-trained vision-language models to derive semantically meaningful skills from offline data, significantly enhancing sample efficiency and adaptability in learning new tasks. Furthermore, the Robust Policy Optimization (RPO) algorithm maintains high policy entropy throughout training, ensuring consistent performance across diverse environments. Collectively, these innovations provide promising solutions to limitations in traditional reinforcement learning approaches [84, 81, 76]. Table 2 presents a comparative analysis of various model-free reinforcement learning methods, illustrating their applications, challenges, and innovations, thereby underscoring their significance in addressing complex real-world problems.

# 8 Conclusion

## 8.1 Ethical and Practical Considerations

Implementing Reinforcement Learning (RL) systems requires careful consideration of ethical and practical dimensions to ensure they align with societal values and function reliably in complex scenarios. Ethical issues, such as the risk of RL systems exploiting reward structures and causing unintended consequences, underscore the necessity for designs that adhere to human values and objectives. Interpretability is crucial for both ethical and practical aspects, as enhancing model transparency aids in debugging and aligning systems with human preferences, thereby fostering trust and accountability. Practically, RL applications must navigate challenges like integrating non-differentiable evaluation metrics typical in logistics and complex domains, which necessitates methods that maintain performance while accommodating these metrics. Tools like ShinRL are essential for understanding and validating the theoretical foundations of RL algorithms, thereby strengthening the field's robustness. An interdisciplinary approach, merging concepts from neuroscience and artificial intelligence, offers promising solutions to these challenges. Advances in neuro-symbolic AI, for instance, provide interpretable models capable of efficiently handling complex tasks, while platforms like Gymnasium facilitate collaborative research and the sharing of best practices. Future research should explore adaptive methods for setting perturbation parameters in Robust Policy Optimization (RPO) and investigate RPO's applicability to diverse RL problems. Enhancements to hybrid approaches, particularly in integrating expert data and online learning, could further improve performance in complex environments. Addressing ethical and practical considerations in RL requires a comprehensive approach that integrates transparency, accountability, and interdisciplinary collaboration, ensuring RL systems are effective and aligned with societal values.

## 8.2 Challenges in Reinforcement Learning

Reinforcement Learning (RL) faces several enduring challenges that impede its widespread application in complex and dynamic environments. A primary challenge is the computational complexity of optimizing RL algorithms, which struggle to efficiently balance reasoning and learning, particularly in large-scale scenarios with high-dimensional state and action spaces. The computational costs associated with optimizing strategies, such as stochastic control barriers (SCBs), may not be sustainable for all applications, complicating the scalability of RL methods. Additionally, the lack of explicit knowledge regarding state transition dynamics complicates the learning process and optimal policy formulation, especially in model-based approaches where inaccuracies in the learned dynamics model can severely degrade performance. Sample inefficiency remains a persistent obstacle, as RL algorithms often require extensive data to learn effective policies, which is particularly problematic in real-world applications like autoscaling. Designing effective reward functions is crucial, as poorly constructed rewards can lead to suboptimal policy learning and unintended behaviors. The adaptation and transferability of learned policies across different tasks and environments present ongoing challenges, particularly in achieving consistent convergence to optimal policies in stochastic settings. Addressing these challenges requires a multifaceted approach that combines algorithmic innovation, computational efficiency, and ethical considerations, paving the way for more effective and reliable RL systems capable of tackling complex decision-making tasks across diverse domains.

### 8.3 Applications and Implications of RL

Reinforcement Learning (RL) has demonstrated significant potential across various domains, offering sophisticated solutions to complex decision-making challenges. In the energy sector, RL has been effectively utilized to optimize building energy management and HVAC systems, leading to substantial energy savings and improved operational performance. In robotics, RL plays a crucial role in developing advanced control policies for tasks like obstacle avoidance and manipulation, increasing the adaptability and precision of autonomous systems. The financial industry benefits from RL through optimized trading strategies and risk management, enhancing decision-making robustness in uncertain market environments. In logistics and supply chain management, RL methods have significantly improved resource allocation and operational efficiency, with practical applications validated in real-world case studies. The implications of RL extend beyond immediate applications, influencing future research directions. Future work may focus on optimizing Transformer architectures for efficiency and integrating dynamic programming techniques for improved performance in complex RL environments. Further research could explore algorithmic robustness and efficiency enhancements, potentially integrating advanced optimization techniques or investigating different task distributions. Despite these advancements, unresolved questions remain regarding the integration of RL with Agent-Based Computational Economics (ACE) and the overall effectiveness of these models in real-world applications. The applications and implications of RL are vast and transformative, with the potential to revolutionize various industries by providing innovative solutions to complex challenges. As research progresses, RL is expected to play an increasingly integral role in shaping the future of artificial intelligence and decision-making technologies.

## References


[1] Ahmad Ahmad, Mehdi Kermanshah, Kevin Leahy, Zachary Serlin, Ho Chit Siu, Makai Mann, Cristian-Ioan Vasile, Roberto Tron, and Calin Belta. Accelerating proximal policy optimization learning using task prediction for solving environments with delayed rewards, 2024.

[2] Mingqi Yuan. Intrinsically-motivated reinforcement learning: A brief introduction, 2022.

[3] Xiujun Li, Lihong Li, Jianfeng Gao, Xiaodong He, Jianshu Chen, Li Deng, and Ji He. Recurrent reinforcement learning: A hybrid approach, 2015.

[4] Thibaut Théate and Damien Ernst. Risk-sensitive policy with distributional reinforcement learning, 2022.

[5] Callum Rhys Tilbury. Reinforcement learning for economic policy: A new frontier?, 2023.

[6] Hitesh Golchha, Sahil Yerawar, Dhruvesh Patel, Soham Dan, and Keerthiram Murugesan. Language guided exploration for rl agents in text environments, 2024.

[7] Alexis Jacq, Johan Ferret, Olivier Pietquin, and Matthieu Geist. Lazy-mdps: Towards interpretable reinforcement learning by learning when to act, 2022.

[8] Brian Tomasik. Do artificial reinforcement-learning agents matter morally?, 2014.

[9] Chao Yu, Shicheng Ye, and Hankz Hankui Zhuo. Reinforcement learning with knowledge representation and reasoning: A brief survey, 2025.

[10] Jack Parker-Holder, Raghu Rajan, Xingyou Song, André Biedenkapp, Yingjie Miao, Theresa Eimer, Baohe Zhang, Vu Nguyen, Roberto Calandra, Aleksandra Faust, et al. Automated reinforcement learning (autorl): A survey and open problems. *Journal of Artificial Intelligence Research*, 74:517–568, 2022.

[11] Vincent Graziano, Faustino Gomez, Mark Ring, and Juergen Schmidhuber. T-learning, 2011.

[12] Yisel Garí, David A. Monge, Elina Pacini, Cristian Mateos, and Carlos García Garino. Reinforcement learning-based application autoscaling in the cloud: A survey, 2020.

[13] Taku Yamagata, Ahmed Khalil, and Raul Santos-Rodriguez. Q-learning decision transformer: Leveraging dynamic programming for conditional sequence modelling in offline rl, 2023.

[14] Kerong Wang, Hanye Zhao, Xufang Luo, Kan Ren, Weinan Zhang, and Dongsheng Li. Bootstrapped transformer for offline reinforcement learning, 2022.

[15] Yu Yang and Pan Xu. Pre-trained language models improve the few-shot prompt ability of decision transformer, 2024.

[16] Mohamed-Amine Chadi and Hajar Mousannif. Understanding reinforcement learning algorithms: The progress from basic q-learning to proximal policy optimization, 2023.

[17] Stefan Werner and Sebastian Peitz. Learning a model is paramount for sample efficiency in reinforcement learning control of pdes, 2023.

[18] Charline Le Lan, Stephen Tu, Mark Rowland, Anna Harutyunyan, Rishabh Agarwal, Marc G. Bellemare, and Will Dabney. Bootstrapped representations in reinforcement learning, 2023.

[19] Borislav Mavrin, Hengshuai Yao, and Linglong Kong. Deep reinforcement learning with decorrelation, 2019.

[20] Budi Kurniawan, Peter Vamplew, Michael Papasimeon, Richard Dazeley, and Cameron Foale. Discrete-to-deep supervised policy learning, 2020.

[21] Ameya Pore, Riccardo Muradore, and Diego Dall'Alba. Dear: Disentangled environment and agent representations for reinforcement learning without reconstruction, 2024.

[22] Liudmyla Nechepurenko, Viktor Voss, and Vyacheslav Gritsenko. Comparing knowledge-based reinforcement learning to neural networks in a strategy game, 2020.

[23] Yunhao Tang and Rémi Munos. Towards a better understanding of representation dynamics under td-learning, 2023.

[24] Jakub Grudzien Kuba, Christian Schroeder de Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation, 2024.

[25] Yihao Wan, Qianwen Xu, and Tomislav Dragičević. Safety-enhanced self-learning for optimal power converter control, 2023.

[26] Juan Sebastian Rojas and Chi-Guhn Lee. Burning red: Unlocking subtask-driven reinforcement learning and risk-awareness in average-reward markov decision processes, 2025.

[27] Jarek Liesen, Chris Lu, Andrei Lupu, Jakob N. Foerster, Henning Sprekeler, and Robert T. Lange. Discovering minimal reinforcement learning environments, 2024.

[28] Yubo Huang, Xuechun Wang, Luobao Zou, Zhiwei Zhuang, and Weidong Zhang. Soft policy optimization using dual-track advantage estimator, 2020.

[29] Lucas N. Alegre, Ana L. C. Bazzan, and Bruno C. da Silva. Optimistic linear support and successor features as a basis for optimal policy transfer, 2022.

[30] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem, 2021.

[31] Daniel Layeghi, Steve Tonneau, and Michael Mistry. Neural lyapunov and optimal control, 2024.

[32] Vivek Veeriah, Tom Zahavy, Matteo Hessel, Zhongwen Xu, Junhyuk Oh, Iurii Kemaev, Hado van Hasselt, David Silver, and Satinder Singh. Discovery of options via meta-learned subgoals, 2021.

[33] Pranay Pasula. Lagrangian duality in reinforcement learning, 2020.

[34] Denis Tarasov, Kirill Brilliantov, and Dmitrii Kharlapenko. Is value functions estimation with classification plug-and-play for offline reinforcement learning?, 2024.

[35] Chuning Zhu, Xinqi Wang, Tyler Han, Simon S. Du, and Abhishek Gupta. Distributional successor features enable zero-shot policy optimization, 2024.

[36] Finn Rietz, Erik Schaffernicht, Stefan Heinrich, and Johannes Andreas Stork. Prioritized soft q-decomposition for lexicographic reinforcement learning, 2024.

[37] Thanh Nguyen-Tang and Raman Arora. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling, and beyond, 2024.

[38] Recovery rl: Safe reinforcement.

[39] Rohan Chitnis, Yingchen Xu, Bobak Hashemi, Lucas Lehnert, Urun Dogan, Zheqing Zhu, and Olivier Delalleau. Iql-td-mpc: Implicit q-learning for hierarchical model predictive control, 2023.

[40] Xiaohan Hu, Yi Ma, Chenjun Xiao, Yan Zheng, and Jianye Hao. Iteratively refined behavior regularization for offline reinforcement learning, 2023.

[41] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

[42] Tao Ma, Xuzhi Yang, and Zoltan Szabo. To switch or not to switch? balanced policy switching in offline reinforcement learning, 2025.

[43] Tianle Zhang, Jiayi Guan, Lin Zhao, Yihang Li, Dongjiang Li, Zecui Zeng, Lei Sun, Yue Chen, Xuelong Wei, Lusong Li, and Xiaodong He. Preferred-action-optimized diffusion policies for offline reinforcement learning, 2024.

[44] Wenxuan Zhou, Steven Bohez, Jan Humplik, Abbas Abdolmaleki, Dushyant Rao, Markus Wulfmeier, Tuomas Haarnoja, and Nicolas Heess. Forgetting and imbalance in robot lifelong learning with off-policy data, 2022.

[45] Zhang-Wei Hong, Pulkit Agrawal, Rémi Tachet des Combes, and Romain Laroche. Harnessing mixed offline reinforcement learning datasets via trajectory weighting, 2023.

[46] Volodymyr Tkachuk, Gellért Weisz, and Csaba Szepesvári. Trajectory data suffices for statistically efficient learning in offline rl with linear $q^\pi$-realizability and concentrability, 2024.

[47] Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, and Chongjie Zhang. Rethinking goal-conditioned supervised learning and its connection to offline rl, 2022.

[48] Tianyu Zheng, Ge Zhang, Xingwei Qu, Ming Kuang, Stephen W. Huang, and Zhaofeng He. More-3s:multimodal-based offline reinforcement learning with shared semantic spaces, 2024.

[49] Seohong Park, Kevin Frans, Sergey Levine, and Aviral Kumar. Is value learning really the main bottleneck in offline rl?, 2024.

[50] Jingxiao Chen, Weiji Xie, Weinan Zhang, Yong yu, and Ying Wen. Offline fictitious self-play for competitive games, 2024.

[51] Jérôme Arjonilla, Abdallah Saffidine, and Tristan Cazenave. Enhancing reinforcement learning through guided search, 2024.

[52] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[53] Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning, 2023.

[54] Bingyi Kang, Xiao Ma, Yirui Wang, Yang Yue, and Shuicheng Yan. Improving and benchmarking offline reinforcement learning algorithms, 2023.

[55] Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Reinforcement learning in dynamic treatment regimes needs critical reexamination, 2024.

[56] David Brandfonbrener, William F. Whitney, Rajesh Ranganath, and Joan Bruna. Offline rl without off-policy evaluation, 2021.

[57] Ziqi Zhang, Xiao Xiong, Zifeng Zhuang, Jinxin Liu, and Donglin Wang. Improving offline-to-online reinforcement learning with q conditioned state entropy exploration, 2024.

[58] Aaron Sonabend-W, Junwei Lu, Leo A. Celi, Tianxi Cai, and Peter Szolovits. Expert-supervised reinforcement learning for offline policy learning and evaluation, 2020.

[59] Jiafei Lyu, Le Wan, Zongqing Lu, and Xiu Li. Off-policy rl algorithms can be sample-efficient for continuous control via sample multiple reuse, 2023.

[60] Eshwar S R, Shishir Kolathaya, and Gugan Thoppe. Improving sample efficiency in evolutionary rl using off-policy ranking, 2023.

[61] Zeyu Zhang, Yi Su, Hui Yuan, Yiran Wu, Rishab Balasubramanian, Qingyun Wu, Huazheng Wang, and Mengdi Wang. Unified off-policy learning to rank: a reinforcement learning perspective, 2023.

[62] JaeYoon Kim, Junyu Xuan, Christy Liang, and Farookh Hussain. A non-monolithic policy approach of offline-to-online reinforcement learning, 2024.

[63] Takayuki Osa and Tatsuya Harada. Discovering multiple solutions from a single task in offline reinforcement learning, 2024.

[64] Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30, 2017.

[65] Joel Oren, Chana Ross, Maksym Lefarov, Felix Richter, Ayal Taitler, Zohar Feldman, Christian Daniel, and Dotan Di Castro. Solo: Search online, learn offline for combinatorial optimization problems, 2021.

[66] Dong Neuck Lee and Michael R. Kosorok. Off-policy reinforcement learning with high dimensional reward, 2024.

[67] Sili Huang, Jifeng Hu, Zhejian Yang, Liwei Yang, Tao Luo, Hechang Chen, Lichao Sun, and Bo Yang. Decision mamba: Reinforcement learning via hybrid selective sequence modeling, 2024.

[68] Jiuqi Wang, Ethan Blaser, Hadi Daneshmand, and Shangtong Zhang. Transformers can learn temporal difference methods for in-context reinforcement learning, 2025.

[69] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization, 2023.

[70] Pouya Hamadanian, Arash Nasr-Esfahany, Malte Schwarzkopf, Siddartha Sen, and Mohammad Alizadeh. Online reinforcement learning in non-stationary context-driven environments, 2024.

[71] Younggyo Seo, Kimin Lee, Ignasi Clavera, Thanard Kurutach, Jinwoo Shin, and Pieter Abbeel. Trajectory-wise multiple choice learning for dynamics generalization in reinforcement learning, 2020.

[72] Amir Mosavi, Pedram Ghamisi, Yaser Faghan, and Puhong Duan. Comprehensive review of deep reinforcement learning methods and applications in economics, 2020.

[73] Sili Huang, Jifeng Hu, Hechang Chen, Lichao Sun, and Bo Yang. In-context decision transformer: Reinforcement learning via hierarchical chain-of-thought, 2024.

[74] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[75] Artemij Amiranashvili, Alexey Dosovitskiy, Vladlen Koltun, and Thomas Brox. Td or not td: Analyzing the role of temporal differencing in deep reinforcement learning, 2018.

[76] Jesse Zhang, Minho Heo, Zuxin Liu, Erdem Biyik, Joseph J Lim, Yao Liu, and Rasool Fakoor. Extract: Efficient policy learning by extracting transferable robot skills from offline data, 2024.

[77] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters in on-policy reinforcement learning? a large-scale empirical study, 2020.

[78] Hongyao Tang and Glen Berseth. Improving deep reinforcement learning by reducing the chain effect of value and policy churn, 2024.

[79] Yunpeng Qing, Shunyu Liu, Jie Song, Huiqiong Wang, and Mingli Song. A survey on explainable reinforcement learning: Concepts, algorithms, challenges, 2023.

[80] Rethinking reinforcement learning for recommendation: A prompt perspective.

[81] Eugene Bykovets, Yannick Metz, Mennatallah El-Assady, Daniel A. Keim, and Joachim M. Buhmann. How to enable uncertainty estimation in proximal policy optimization, 2022.

[82] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. The potential of the return distribution for exploration in rl, 2018.

[83] What matters in on-policy reinfo.

[84] Robust policy optimization in deep reinforcement learning.

[85] Xiaowei Mao, Haomin Wen, Hengrui Zhang, Huaiyu Wan, Lixia Wu, Jianbin Zheng, Haoyuan Hu, and Youfang Lin. Drl4route: A deep reinforcement learning framework for pick-up and delivery route prediction, 2023.

[86] Qianli Shen, Yan Li, Haoming Jiang, Zhaoran Wang, and Tuo Zhao. Deep reinforcement learning with robust and smooth policy, 2020.

[87] Shiron Thalagala, Pak Kin Wong, and Xiaozheng Wang. Broad critic deep actor reinforcement learning for continuous control, 2024.

[88] Vektor Dewanto and Marcus Gallagher. Examining average and discounted reward optimality criteria in reinforcement learning, 2022.

[89] Henrique Donâncio, Antoine Barrier, Leah F. South, and Florence Forbes. Dynamic learning rate for deep reinforcement learning: A bandit approach, 2025.

[90] Chris Reinke. Time adaptive reinforcement learning, 2020.

[91] Mathukumalli Vidyasagar. A tutorial introduction to reinforcement learning, 2023.

[92] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[93] Emre O Neftci and Bruno B Averbeck. Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, 1(3):133–143, 2019.

[94] Kewen Ding. Addressing the issue of stochastic environments and local decision-making in multi-objective reinforcement learning, 2022.

[95] Hardik Meisheri, Vinita Baniwal, Nazneen N Sultana, Balaraman Ravindran, and Harshad Khadilkar. Reinforcement learning for multi-objective optimization of online decisions in high-dimensional systems, 2019.

[96] Elad Sarafian, Aviv Tamar, and Sarit Kraus. Constrained policy improvement for safe and efficient reinforcement learning, 2019.

[97] Matteo Turchetta, Andreas Krause, and Sebastian Trimpe. Robust model-free reinforcement learning with multi-objective bayesian optimization, 2019.

[98] Wesley Cowan, Michael N. Katehakis, and Daniel Pirutinsky. Reinforcement learning: a comparison of ucb versus alternative adaptive policies, 2019.

[99] Chris Lu, Jakub Grudzien Kuba, Alistair Letcher, Luke Metz, Christian Schroeder de Witt, and Jakob Foerster. Discovered policy optimisation, 2022.

[100] Han Qi, Yi Su, Aviral Kumar, and Sergey Levine. Data-driven offline decision-making via invariant representation learning, 2022.

[101] Isaac Grosof, Siva Theja Maguluri, and R. Srikant. Convergence for natural policy gradient on infinite-state average-reward markov decision processes, 2024.

[102] Chenlu Ye, Rui Yang, Quanquan Gu, and Tong Zhang. Corruption-robust offline reinforcement learning with general function approximation, 2024.

[103] Zhenyang Shi and Surya P. N. Singh. Soft actor-critic with cross-entropy policy optimization, 2021.

[104] Raphael Fonteneau and L. A. Prashanth. Simultaneous perturbation algorithms for batch off-policy search, 2014.

[105] Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning, 2021.

[106] Aaqib Parvez Mohammed and Matias Valdenegro-Toro. Can reinforcement learning for continuous control generalize across physics engines?, 2020.

[107] Léonard Hussenot, Robert Dadashi, Matthieu Geist, and Olivier Pietquin. Show me the way: Intrinsic motivation from demonstrations, 2021.

[108] Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Başar. Model-free non-stationary rl: Near-optimal regret and applications in multi-agent rl and inventory control, 2022.

[109] Arthur Argenson and Gabriel Dulac-Arnold. Model-based offline planning, 2021.

[110] Safa Alver and Doina Precup. A look at value-based decision-time vs. background planning methods across different settings, 2024.

[111] Giacomo Arcieri, David Wölfle, and Eleni Chatzi. Which model to trust: Assessing the influence of models on the performance of reinforcement learning algorithms for continuous control tasks, 2022.

[112] Mohamad Fares El Hajj Chehade, Young ho Cho, Sandeep Chinchali, and Hao Zhu. Should we use model-free or model-based control? a case study of battery management systems, 2024.

[113] Juntao Ren, Gokul Swamy, Zhiwei Steven Wu, J. Andrew Bagnell, and Sanjiban Choudhury. Hybrid inverse reinforcement learning, 2024.

**Disclaimer:**

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.