Online Reinforcement Learning for Decision-Making and Explicit Policy Learning in Autonomous Agents: A Systematic Literature Review

Course of Intelligent Systems Engineering

Pablo Sebastian Vargas Grateron
ID: 0001137787
email: pablo.vargasgrateron@studio.unibo.it

18 January 2026

Abstract

Online reinforcement learning (RL) enables autonomous agents to adapt their policies in real time, responding to dynamic and uncertain environments. This systematic literature review (SLR) examines high-impact studies from 2004 to 2026 on online RL methods for decision-making and explicit policy learning in autonomous agents, excluding approaches that do not adapt during execution. We analyze algorithmic strategies for handling non-stationarity, exploration-exploitation trade-offs, and multi-agent interactions. The review highlights emerging trends, practical implementations, and remaining challenges, offering actionable insights for designing adaptive and efficient online RL controllers in robotics, UAVs, and intelligent systems.

Contents

List of Figures

Chapter 1

Introduction

Reinforcement Learning (RL) has become a central tool for developing autonomous agents capable of making complex decisions in dynamic and uncertain environments. Traditionally, many RL approaches have been designed for offline or simulated settings, requiring pretraining and not updating the policy during execution. However, in real-world applications—such as robotics, autonomous vehicles, multi-agent systems, and drones—agents must continuously adapt to changing environmental conditions, interactions with other agents, and evolving operational constraints.

Online reinforcement learning (RL) refers to a class of algorithms in which agents learn and adapt their policies through continuous interaction with the environment, updating their behavior in real-time as new data becomes available. Unlike offline or batch RL, which relies on pre-collected datasets, online methods enable agents to respond dynamically to changes in the environment, making them particularly suitable for autonomous systems operating in uncertain or non-stationary settings. Policy learning in this context involves optimizing a mapping from states to actions, allowing agents to make decisions that maximize long-term rewards. The ability to explicitly adapt policies during execution distinguishes online RL from static approaches, providing both flexibility and robustness in complex decision-making tasks, from robotic control to traffic management and multi-agent coordination.

In this context, online reinforcement learning emerged as a critical approach, enabling agents to learn and update their policies in real time while handling uncertainty and environmental complexity. This systematic literature review (SLR) focuses on online RL methods that support active policy adaptation, excluding approaches that do not modify behavior during execution. By analyzing high-impact studies published from 2004 to 2026, this work identifies emerging trends, algorithmic strategies, application domains, and open challenges, aiming to provide guidance for developing more adaptive and efficient autonomous agents.

Chapter 2

Methodology

In this chapter, we outline the methodology adopted for conducting the systematic literature review (SLR) on online reinforcement learning methods for decision-making and explicit policy learning in autonomous agents. The process involves a structured selection of relevant studies, extraction of key information, and synthesis of findings to identify trends, challenges, and gaps in the existing literature. We describe the inclusion and exclusion criteria, the databases and search queries used, and the approach for analyzing the selected studies.

2.1 Research Questions and Objectives

The objectives of this Systematic Literature Review (SLR) are the following:

- To identify reinforcement learning approaches that update policies during execution and can be considered truly online, clearly distinguishing them from offline or static methods.
- To analyze how autonomous agents address environmental non-stationarity and the presence of multiple agents or complex variables during the online decision-making process.
- To map the main application domains (e.g., robotics, UAVs, smart grids, autonomous vehicles) and to evaluate the effectiveness of the proposed methods with respect to key metrics such as convergence, adaptability, and robustness.

Based on these objectives, the following research questions have been formulated:

- **Question 1:** Which online reinforcement learning (RL) methods have been proposed to enable dynamic policy adaptation in autonomous agents during execution?

- **Question 2:** Which strategies are adopted to balance exploration and exploitation in dynamic and non-stationary environments for online learning?
- **Question 3:** What are the main applications and reported performance of online RL methods in real-world or simulated autonomous agent scenarios?

2.2 Search Strategy

The literature search was conducted using **Scopus**, selected for its broad coverage of peer-reviewed publications in computer science and engineering, as well as its advanced support for Boolean logic and proximity operators. To ensure relevance and consistency, the search was restricted to the `TITLE-ABS-KEY` fields and limited to publications classified under the Computer Science and Engineering subject areas.

The search query was structured to capture studies focusing on online reinforcement learning and related paradigms that support continuous or incremental policy adaptation. Approaches based on offline or batch learning, imitation-based methods, and off-policy evaluation were explicitly excluded to ensure that only genuinely online policy adaptation methods were retained.

The following search string was used as the exact query input in the Scopus database:

```
 TITLE-ABS-KEY ( ( "online reinforcement learning" OR "continual reinforcement learning"
OR "lifelong reinforcement learning" OR ( reinforcement W/2 learning W/3 ( online OR continual
OR lifelong OR incremental ) ) OR "non-stationary" OR nonstationary OR "concept drift"
OR "changing environment*" ) AND ( "policy gradient" OR "actor-critic" OR "policy improvement"
OR "stochastic polic*" OR "explicit polic*" ) AND ( robot* OR "autonomous agent*" OR "autonomous
robot*" OR "mobile robot*" OR UAV OR drone* OR "self-driving" OR "multi-agent" ) ) AND
NOT TITLE-ABS-KEY ( "offline reinforcement learning" OR "batch reinforcement learning"
OR "behavioral cloning" OR "imitation learning" OR "inverse reinforcement learning" OR
"off-policy evaluation" ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA
, "ENGI" ) )
```

The keywords used in the query were organized into three conceptual groups: *learning paradigm*, *policy representation*, and *application domain*.

The learning paradigm group included terms such as:

- *online reinforcement learning*

- *continual reinforcement learning*
- *lifelong reinforcement learning*
- *incremental learning*
- *non-stationary environments*
- *concept drift*

The policy representation group focused on explicit policy-based methods, incorporating keywords related to:

- *policy gradient*
- *actor-critic*
- *policy improvement*
- *stochastic or explicit policies*

Finally, the application domain group constrained the search to autonomous and robotic systems, including:

- *autonomous agents*
- *mobile robots*
- *UAVs*
- *self-driving systems*
- *multi-agent environments*

2.3 Study Selection

The study selection process was designed to systematically identify primary studies relevant to the objectives of this systematic literature review while minimizing selection bias. The selection procedure was conducted in multiple stages, each progressively refining the pool of candidate studies.

1. The **first stage** involved an initial screening of the retrieved publications based on their title, abstract, and bibliographic metadata. During this phase, studies that clearly failed to satisfy the predefined inclusion criteria were excluded.

2. The **second stage** consisted of a full-text assessment of all studies that passed the initial screening. Each article was examined in detail to evaluate its methodological relevance, with particular attention to whether the proposed approaches addressed the objectives of this review. Studies lacking sufficient methodological detail or falling outside the defined scope were excluded at this stage.

3. The **final stage** resulted in the identification of the primary studies included in the qualitative synthesis. These studies constitute the foundation for the subsequent analysis and classification, enabling a structured comparison of online reinforcement learning methods across different learning paradigms, adaptation strategies, and application domains.

2.3.1 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria were defined to ensure that only studies relevant to the objectives of this systematic literature review were considered.

Inclusion Criteria

- Studies proposing or analyzing reinforcement learning methods that update policies *online* during execution.
- Approaches explicitly based on policy learning, such as policy gradient or actor-critic methods.
- Studies addressing decision-making in dynamic or non-stationary environments.
- Applications involving autonomous agents, robotic systems, UAVs, self-driving vehicles, or multi-agent systems.
- Publications written in English.
- Studies published between 2023 and 2026. For publications prior to 2023, only studies with more than 50 citations were considered.

Exclusion Criteria

- Studies focused exclusively on offline or batch reinforcement learning.
- Approaches based on imitation learning, behavioral cloning, or inverse reinforcement learning.
- Works that do not involve explicit policy adaptation during execution.

- Studies lacking sufficient methodological detail to support qualitative analysis.
- Non-peer-reviewed material, including surveys, tutorials, theses, posters, or gray literature.
- Publications not related to autonomous or robotic decision-making.
- Studies and publications that are not accessible through institutional or educational accounts, or via open access sources.

2.4 Data Extraction and Synthesis

This section describes the procedures adopted to extract relevant information from the selected studies and to synthesize the collected data in a structured and systematic manner.

2.4.1 Data Extraction

A structured data extraction process was applied to all primary studies included in the final selection. For each study, a predefined set of attributes was extracted to ensure consistency and comparability across the literature. The extracted data included the reinforcement learning paradigm adopted, the policy representation and learning strategy, the type of environment considered (stationary or non-stationary), and the application domain.

2.4.2 Data Synthesis

The data synthesis was conducted using a qualitative approach aimed at identifying common patterns and differences across the selected studies. The extracted information was analyzed to classify the literature according to learning paradigms, policy adaptation strategies, and application domains. This classification enabled a systematic comparison of online reinforcement learning methods with respect to their adaptability, robustness, and suitability for autonomous decision-making in dynamic environments.

Rather than aggregating quantitative results, the synthesis focused on interpreting methodological trends and conceptual similarities among studies. The outcomes of this synthesis were used to identify prevailing research directions, limitations of existing approaches, and open challenges that warrant further investigation.

Chapter 3

Results

3.1 Q1: Which online reinforcement learning (RL) methods have been proposed to enable dynamic policy adaptation in autonomous agents during execution?

This section focuses on online RL methods that enable policy adaptation during execution, independently of the specific exploration–exploitation strategies adopted, which are discussed separately in Q2.

3.1.1 Online Actor-Critic Based Methods

Online actor-critic methods represent the most prevalent class of approaches for enabling dynamic policy adaptation during execution. In these methods, the control policy (actor) and the value or action-value function (critic) are learned and updated concurrently through continuous interaction with the environment. The actor generates actions according to the current policy, while the critic evaluates these actions using temporal-difference errors or Bellman residuals computed from online experience. This tight coupling enables stable learning in continuous state and action spaces, making actor-critic architectures particularly suitable for real-time control and autonomous decision-making [14, 22, 15, 5, 27, 9, 2, 3, 6].

Across the analyzed literature, several **variations of online actor-critic learning** are proposed, which can be grouped as follows:

- **Adaptive and identifier-based actor-critic methods** (Example: Fig. 3.1), where classical structures are augmented with identifier networks or adaptive components to cope with unknown or nonlinear system dynamics. Policy parameters are updated online using gradient-based rules driven by real-time measurements [14, 15].

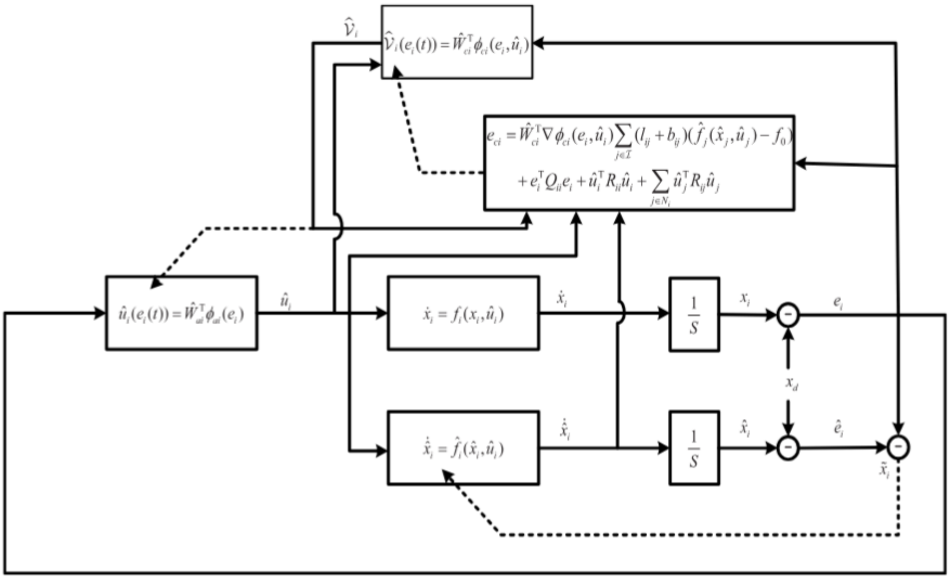

$\hat{\mathcal{V}}_i$
$\hat{\mathcal{V}}_i(e_i(t)) = \hat{W}_{ci}^{\mathrm{T}}\phi_{ci}(e_i,\hat{u}_i)$
$e_{ci} = \hat{W}_{ci}^{\mathrm{T}}\nabla\phi_{ci}(e_i,\hat{u}_i)\sum_{j\in\mathcal{I}}(l_{ij}+b_{ij})(\hat{f}_j(\hat{x}_j,\hat{u}_j)-f_0)$
$+e_i^{\mathrm{T}}Q_{ii}e_i+\hat{u}_i^{\mathrm{T}}R_{ii}\hat{u}_i+\sum_{j\in N_i}\hat{u}_j^{\mathrm{T}}R_{ij}\hat{u}_j$
$\hat{u}_i(e_i(t)) = \hat{W}_{ai}^{\mathrm{T}}\phi_{ai}(e_i)$
$\hat{u}_i$
$\dot{x}_i = f_i(x_i,\hat{u}_i)$
$\dot{x}_i$
$\frac{1}{S}$
x_i
e_i
x_d
$\dot{\hat{x}}_i = \hat{f}_i(\hat{x}_i,\hat{u}_i)$
$\dot{\hat{x}}_i$
$\frac{1}{S}$
$\hat{x}_i$
$\hat{e}_i$
$\tilde{x}_i$

Figure 3.1: Actor–critic–identifier architecture, reproduced from [14].

- **Deep actor-critic variants** (Example: Fig. 3.2), such as Deep Deterministic Policy Gradient (DDPG) and other widely used continuous-control methods, are commonly adopted in the literature for high-dimensional continuous control problems. Within the surveyed works, DDPG-based formulations are explicitly applied to UAV navigation and multi-agent coordination tasks [26, 27], while other studies adopt deep actor-critic architectures for continuous control without relying on a specific DDPG formulation [5, 9, 8, 11, 6].

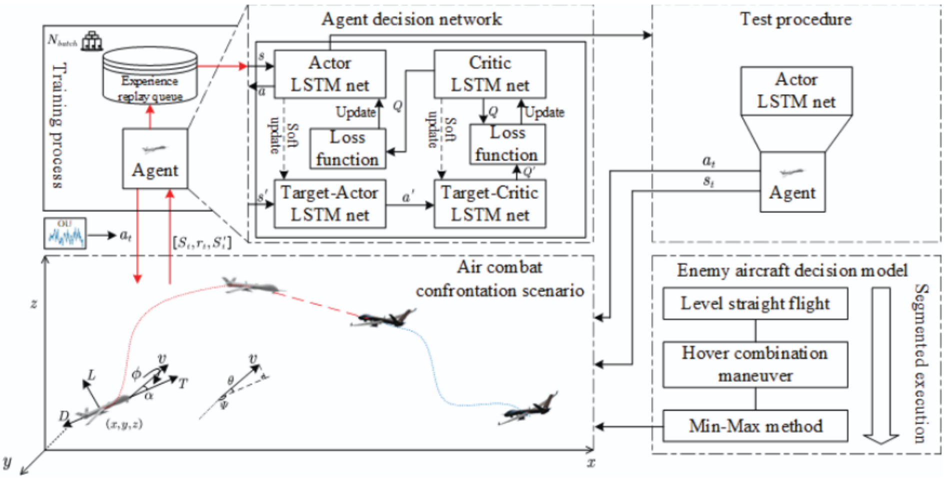


Figure 3.2: A learning framework for LSTM-DDPG maneuver decision making algorithm for close air combat, reproduced from [27].

Several works further extend online actor-critic methods by explicitly addressing **stability, safety, and constraint handling**:

- **Safety- and stability-aware actor-critic formulations**, in which the critic incorporates Lyapunov-based conditions, barrier functions, or admissibility constraints to ensure that online policy updates do not violate safety or stability requirements [15, 2, 3, 10].
- **Asynchronous and decentralized actor-critic methods**, particularly in multi-agent settings, where each agent updates its policy online using local observations while mitigating the non-stationarity induced by other learning agents [19, 26, 25].

Additional extensions of the actor-critic framework include:

- **Curriculum-based, hierarchical, and skill-oriented actor-critic approaches**, where the policy adapts online at multiple levels of abstraction or decision-making [26, 8].
- **Energy-aware and resource-constrained actor-critic methods**, in which the reward function and critic explicitly account for energy consumption, battery constraints, or communication costs, enabling adaptive trade-offs during execution [30, 4, 6].

3.1.2 Model-Free Online Adaptive Control via Reinforcement Learning

Model-free online adaptive control via reinforcement learning represents a prominent class of approaches aimed at enabling autonomous agents to adapt their control policies during execution without relying on explicit models of the system dynamics [5, 7, 25]. In these methods, the agent interacts directly with the environment and incrementally updates its policy or value function based on observed state-action-reward transitions. By avoiding the identification of system equations or accurate physical models, model-free online RL is particularly well suited to complex, uncertain, or time-varying control scenarios [28].

Across the analyzed literature, model-free online adaptive control is typically instantiated through the following learning paradigms:

- **Online value-based methods**, including value iteration and Q-learning formulations, where action-value functions are updated online during task execution [10, 17, 18].
- **Policy-gradient-based approaches**, in which policy parameters are adjusted directly through gradient estimates computed from real-time interaction data [17].

Several studies further apply model-free online reinforcement learning to **distributed and multi-agent control problems**, such as consensus, tracking, and coordination tasks, where centralized modeling and control are impractical or infeasible [7, 25, 10]. Overall, model-free online adaptive control provides a robust and flexible foundation for decision-making in settings where accurate modeling is difficult and continuous online adaptation is essential.

3.1.3 Continual and Lifelong Online Reinforcement Learning

Continual and lifelong online reinforcement learning methods aim to enable autonomous agents to adapt their policies over long time horizons while interacting with a sequence of tasks or nonstationary environments [16, 2, 3, 4]. Unlike standard online reinforcement learning approaches that primarily focus on short-term adaptation, these methods explicitly address the challenge of maintaining and reusing previously acquired knowledge while learning from new experiences during execution, thereby avoiding catastrophic forgetting and repeated retraining.

Across the analyzed literature, continual and lifelong online RL approaches can be characterized according to the following key mechanisms:

- **Knowledge retention and regularization-based methods**, in which policy updates are constrained to preserve important parameters or representations learned in previous tasks or environments. These approaches allow agents to adapt online while preventing abrupt degradation of previously acquired skills [2, 3].
- **Incremental and lifelong policy learning strategies**, where agents continuously refine their policies as new tasks or environmental conditions are encountered, without resetting the learning process or retraining from scratch. Such methods emphasize smooth policy evolution over extended deployments [16, 4].

Several works apply continual and lifelong online reinforcement learning to **robotic and multi-agent control scenarios**, including formation control, tracking, and energy-aware coordination, where agents must operate continuously under evolving dynamics, constraints, or mission objectives [2, 3, 4]. In these settings, online policy updates are combined with mechanisms for preserving previously learned behaviors, ensuring stability and sustained performance over

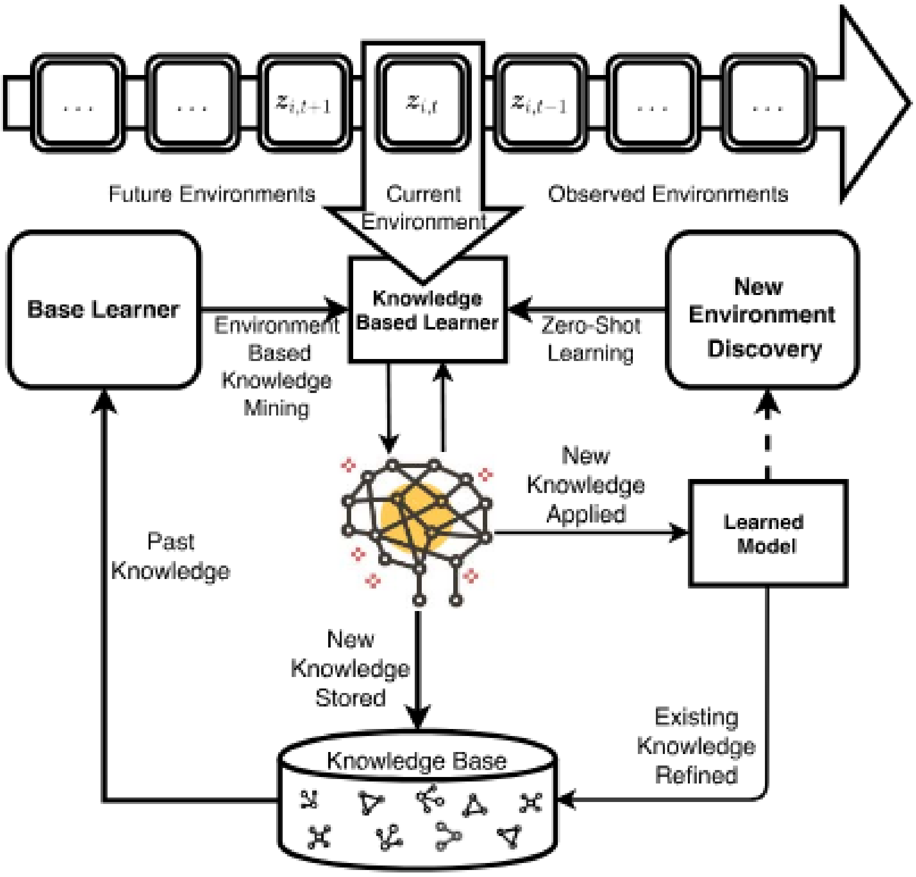


Figure 3.3: Flow diagram of the proposed lifelong RL algorithm, reproduced from [4].

time.

Overall, the surveyed studies demonstrate that continual and lifelong online reinforcement learning provides an effective framework for long-term autonomous operation in nonstationary environments. By combining online adaptation with mechanisms for knowledge preservation and transfer, these approaches enable agents to sustain performance across extended deployments, making them especially suitable for real-world autonomous systems that must learn continuously over time.

3.1.4 Offline-to-Online Policy Adaptation

Offline-to-online policy adaptation methods combine offline training with online reinforcement learning to enable dynamic policy refinement during execution [13, 30, 6]. In these approaches, an initial policy is learned offline using simulation data, historical datasets, or simplified environment models, and is subsequently deployed and adapted online through direct interaction with the real environment. This hybrid learning paradigm reduces exploration costs and safety risks during deployment, while preserving the ability to adapt to nonstationary dynamics and modeling inaccuracies.

Across the analyzed literature, offline-to-online policy adaptation can be characterized according to the following mechanisms:

- **Offline pre-training with online fine-tuning**, where a policy trained offline is incrementally refined online using reinforcement learning updates driven by real-time interaction data. This approach allows agents to compensate for discrepancies between simulated and real-world environments [30, 6].
- **Hybrid learning pipelines**, in which offline learning provides an initial policy prior, while online adaptation is explicitly designed to handle changing environmental conditions, resource constraints, or unforeseen disturbances encountered during execution [13].

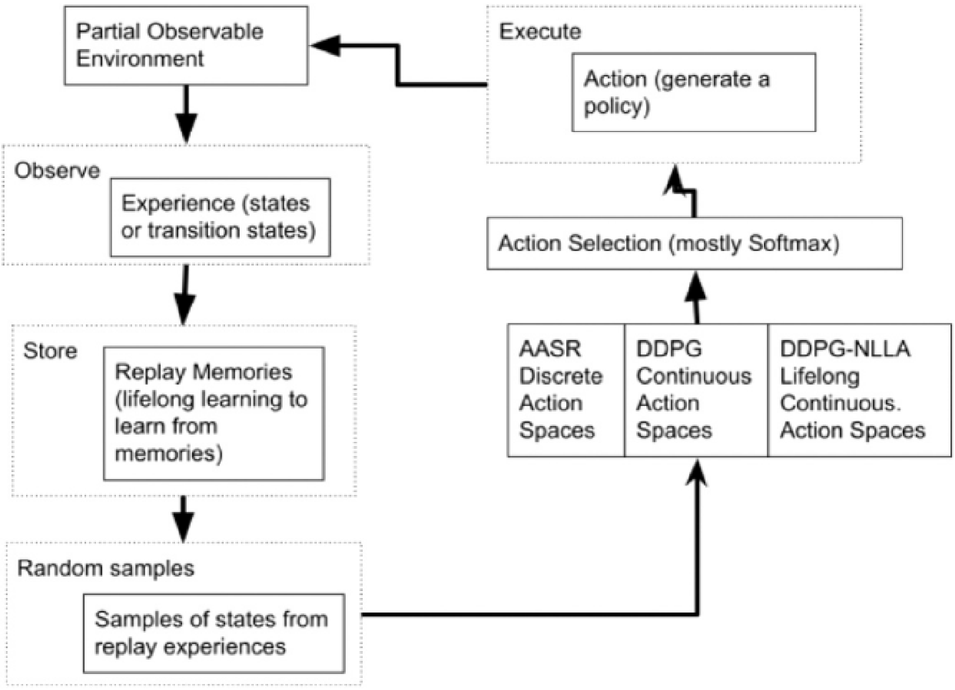

Partial Observable Environment
Execute
Action (generate a policy)
Observe
Experience (states or transition states)
Action Selection (mostly Softmax)
Store
Replay Memories (lifelong learning to learn from memories)
AASR Discrete Action Spaces
DDPG Continuous Action Spaces
DDPG-NLLA Lifelong Continuous. Action Spaces
Random samples
Samples of states from replay experiences

Figure 3.4: Proposed Deep Asynchronous Autonomous Learning System, reproduced from [13].

Several studies apply offline-to-online policy adaptation to **autonomous systems operating under physical and resource constraints**, such as UAV navigation, robotic control, and energy-aware decision-making [30, 6]. In these settings, online refinement plays a critical role in bridging the gap between simulation and real-world deployment, enabling agents to adjust their behavior in response to sensor noise, unmodeled dynamics, and time-varying operational conditions.

Overall, the surveyed works indicate that offline-to-online policy adaptation represents a pragmatic compromise between purely offline and purely online learning. By combining reliable initial behavior with continuous online adaptation, these methods effectively balance learning efficiency, safety, and adaptability, making them well suited for real-world autonomous agents that must operate robustly in dynamic and uncertain environments.

3.1.5 Uncertainty-Aware Online Policy Adaptation

Uncertainty-aware online policy adaptation methods explicitly incorporate measures of uncertainty into the reinforcement learning process in order to improve robustness and adaptability during execution [31]. In these approaches, uncertainty estimates related to the environment, the policy, or the value function are integrated into the online decision-making process, allowing agents to adjust their behavior dynamically in response to incomplete information, nonstationary dynamics, or unpredictable interactions.

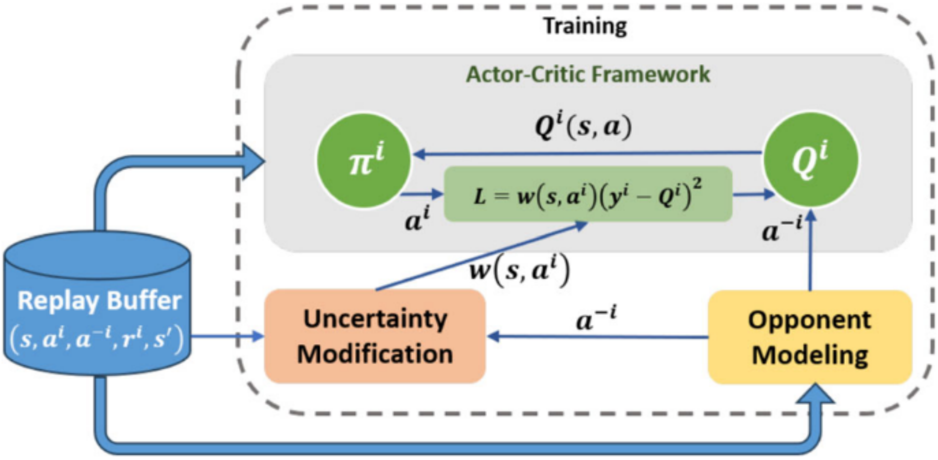


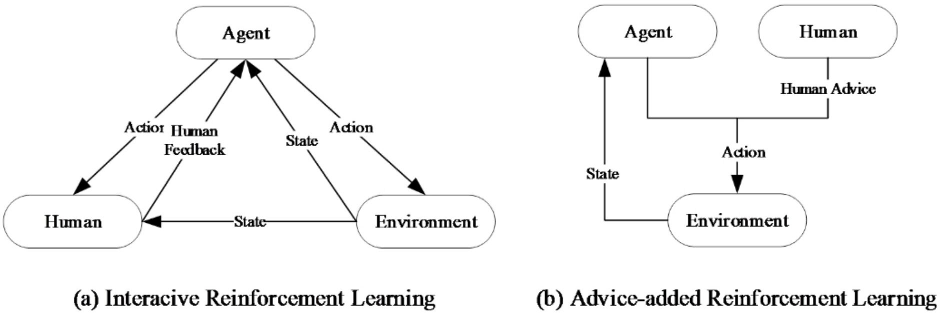
Figure 3.5: Online reinforcement learning framework, reproduced from [31].

Such methods are particularly relevant in dynamic and multi-agent environments, where uncertainty may arise from partial observability, stochastic interactions, or the concurrent learning of other agents. By accounting for uncertainty during execution, agents can react more cautiously when confidence in current policy estimates decreases, or adapt their behavior to mitigate the risks associated with unreliable predictions. This leads to improved stability and performance compared to uncertainty-agnostic online learning approaches.

Overall, the analyzed study indicates that incorporating uncertainty into online reinforcement learning provides an effective mechanism for enhancing adaptability and robustness. By explicitly leveraging uncertainty estimates during policy adaptation, uncertainty-aware approaches enable autonomous agents to maintain reliable performance in complex and evolving environments, where purely deterministic or confidence-agnostic policies may be insufficient.

3.1.6 Knowledge-Integrated Online Reinforcement Learning

Knowledge-integrated online reinforcement learning methods enhance online policy adaptation by incorporating prior knowledge, such as heuristic rules or domain constraints, into the learning process [21, 16]. By guiding exploration and policy updates, these approaches improve learning efficiency and stability, particularly in complex or safety-critical environments.



(a) Interacive Reinforcement Learning

(b) Advice-added Reinforcement Learning

Figure 3.6: The related work of RL integrating prior knowledge, reproduced from [21].

In the surveyed literature, prior knowledge is used to complement online reinforcement learning rather than replace it. Knowledge integration is typically realized by constraining or shaping the action space, modifying reward functions, or biasing policy updates using task-specific or domain-informed information. This allows agents to exploit existing insights while retaining the ability to adapt dynamically to new or changing conditions, reducing unsafe or inefficient exploration.

Several studies apply knowledge-integrated online RL to multi-agent and navigation problems, showing that the inclusion of prior knowledge can accelerate convergence and improve stability during online adaptation [21, 16]. In these scenarios, agents leverage shared or task-specific knowledge to coordinate their behavior more effectively, while still updating their policies online based on real-time interaction data.

Overall, the analyzed works indicate that knowledge-integrated online reinforcement learning provides a practical and efficient mechanism for dynamic policy adaptation. By balancing prior guidance with online learning, these methods enable autonomous agents to adapt reliably in complex and evolving environments while maintaining interpretable and stable behavior.

3.1.7 Skill-Conditioned and Hierarchical Online Reinforcement Learning

Skill-conditioned and hierarchical online reinforcement learning methods introduce structured decision-making by decomposing complex tasks into reusable skills or sub-policies, enabling efficient online adaptation at multiple levels of abstraction [26, 8]. By organizing decision-making hierarchically, these approaches reduce the complexity of online learning while promoting knowledge reuse across tasks and environments.

Within the surveyed literature, skill-conditioned and hierarchical online RL is commonly implemented by conditioning policies on latent skill variables or by adopting multi-level architectures in which high-level policies govern skill selection, while low-level controllers execute primitive actions. Online adaptation typically occurs at the strategic level of the hierarchy, allowing agents to adjust their behavior in response to environmental changes without requiring full relearning of low-level control policies.

Several studies apply these methods to navigation, multi-agent coordination, and curriculum-based learning scenarios, where agents must handle tasks of increasing complexity or switch between distinct behavioral modes [26, 8]. In these settings, hierarchical and skill-based representations enable rapid adaptation to new situations while preserving previously acquired competencies.

Overall, the analyzed works indicate that skill-conditioned and hierarchical online reinforcement learning provides a scalable framework for dynamic policy adaptation. By combining online learning with structured policy representations, these approaches enable autonomous agents to adapt efficiently and robustly in complex environments, while maintaining interpretability and modularity in the learning process.

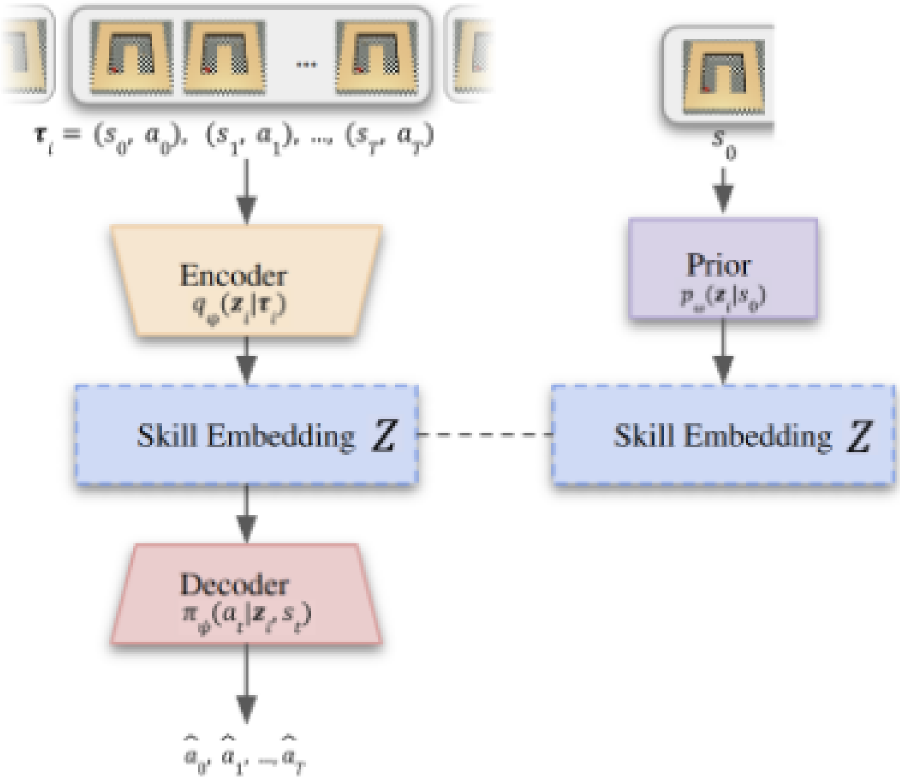

$\tau_i = (s_0, a_0), (s_1, a_1), ..., (s_T, a_T)$
Encoder
$q_\varphi(z_i|\tau_i)$
Skill Embedding Z
Decoder
$\pi_\varphi(a_t|z_t, s_t)$
$\hat{a}_0, \hat{a}_1, ..., \hat{a}_T$
s_0
Prior
$p_\omega(z_i|s_0)$
Skill Embedding Z

Figure 3.7: Overview of the autoencoder model designed for the simultaneous learning of skill embeddings and skill priors, reproduced from [8].

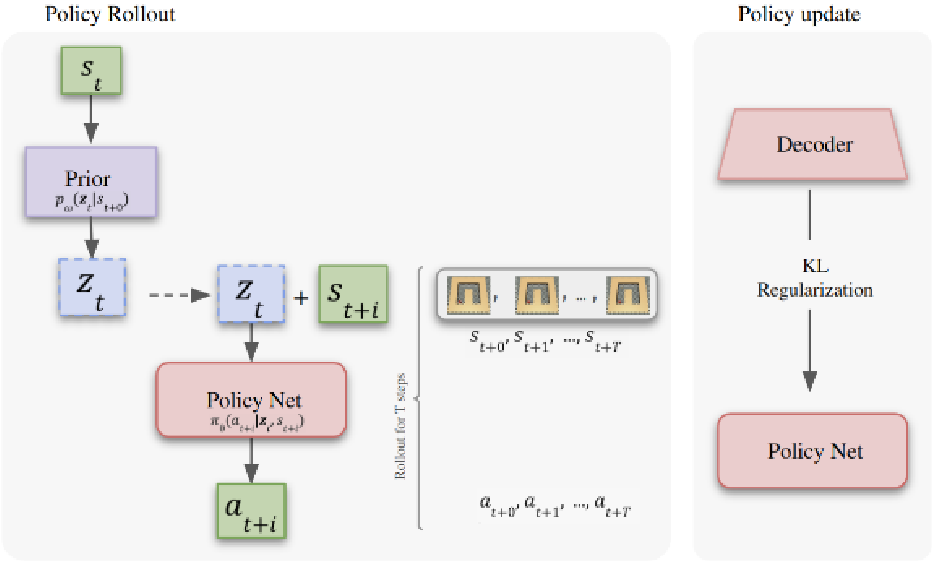


Figure 3.8: Overview of the training process of SCORE, reproduced from [8].

3.1.8 Online Multi-Agent Reinforcement Learning

Online multi-agent reinforcement learning (MARL) methods enable multiple agents to adapt their policies during execution while interacting in shared environments [21, 14, 22, 19, 26, 7, 29, 25, 10, 31]. In contrast to single-agent online RL, MARL introduces additional challenges due to the nonstationarity induced by concurrently learning agents, partial observability, and the need for coordination or competition.

In the surveyed literature, online MARL approaches are predominantly implemented through decentralized or partially centralized learning schemes, where agents perform online policy updates using local information or limited shared signals. Asynchronous update mechanisms are frequently employed to mitigate instability caused by simultaneous learning, allowing agents to adapt at different time scales during execution. These strategies improve scalability and robustness, particularly in large-scale or distributed systems.

Several studies apply online MARL to cooperative control, consensus, scheduling, formation control, and coordination tasks, demonstrating that online adaptation enables agents to respond effectively to dynamic environments and evolving group behaviors [21, 7, 25, 10]. In addition, scalable online reinforcement learning approaches have been proposed for large-scale multi-component systems, where a centralized or aggregated decision-making entity adapts online to changing system requirements [12].

Overall, the analyzed works indicate that online multi-agent reinforcement learning provides a flexible and powerful framework for distributed decision-making in dynamic environments. By enabling agents to adapt their policies online while explicitly accounting for interactions with other learning agents, MARL approaches support scalable and resilient autonomous systems capable of long-term operation under nonstationary conditions.

3.1.9 Real-Time / Pure Online Reinforcement Learning

Real-time or pure online reinforcement learning methods perform learning entirely or predominantly during execution, without relying on extensive offline training [20, 23, 6]. In these approaches, policies are initialized with minimal prior knowledge and continuously updated through real-time interaction with the environment, demonstrating the feasibility of online adaptation under strict timing and computational constraints.

In the surveyed literature, real-time online RL is primarily applied to scenarios where accurate system models or large offline datasets are unavailable, or where environmental conditions evolve too rapidly to be captured through offline training. Policy updates are performed incrementally using real-time feedback, prioritizing learning stability, sample efficiency, and low-latency updates to ensure that adaptation does not interfere with ongoing operation.

Several studies demonstrate the applicability of real-time online RL in physical systems such as robots and unmanned aerial vehicles, where agents must adapt their behavior while interacting with the real world [20, 23, 6]. In these settings, online learning enables agents to compensate for unmodeled dynamics, sensor noise, and environmental disturbances as they occur, without interrupting execution or requiring offline retraining.

Overall, the analyzed works indicate that real-time and pure online reinforcement learning provides a powerful mechanism for achieving high adaptability in dynamic and uncertain environments. By learning exclusively from online interaction, these methods enable autonomous agents to operate continuously and autonomously, making them particularly suitable for real-world applications where immediacy and robustness are critical [28].

3.2 Q2: Which strategies are adopted to balance exploration and exploitation in dynamic and non-stationary environments for online learning?

This section analyzes the strategies adopted in the literature to regulate the trade-off between exploration and exploitation in online reinforcement learning, with particular emphasis on dynamic and non-stationary environments.

3.2.1 Adaptive Exploration Scheduling

Several studies adopt adaptive exploration scheduling strategies to balance exploration and exploitation during online learning. Instead of relying on fixed exploration parameters, these approaches dynamically adjust exploration intensity based on learning progress, performance variations, or detected changes in the environment [2, 3, 7, 25]. By modulating exploration online, agents can increase exploratory behavior when performance degrades or environmental conditions change, while gradually favoring exploitation once stable behavior is recovered. Such adaptive scheduling mechanisms are particularly effective in non-stationary settings, where static exploration policies may lead to premature convergence or degraded performance.

3.2.2 Uncertainty-Driven Exploration

Uncertainty-aware exploration strategies explicitly regulate the exploration or exploitation trade-off based on estimates of uncertainty in the policy, value function, or environment dynamics. In these approaches, exploration is encouraged in regions of high uncertainty and reduced as confidence in the learned policy increases [31, 21]. This mechanism allows agents to respond effectively to non-stationary conditions and unpredictable interactions, particularly in multi-agent environments where uncertainty arises from concurrent learning processes. By incorporating uncertainty into action selection or policy updates, these methods improve robustness and reduce the risk of overconfident exploitation in poorly understood regions of the state-action space.

3.2.3 Continual and Lifelong Exploration Strategies

In continual and lifelong online reinforcement learning, exploration strategies are designed to prevent premature convergence while enabling long-term adaptation across changing tasks or environments. Rather than diminishing exploration permanently, these approaches maintain or reactivate exploratory behavior when new conditions are encountered [2, 3, 4, 16]. Such mechanisms mitigate catastrophic forgetting and allow agents to refine previously learned policies while acquiring new knowledge. By sustaining controlled exploration over extended deployments, lifelong learning approaches support stable performance in non-stationary and evolving environments.

3.2.4 Exploration Strategies in Multi-Agent and Non-Stationary Settings

In multi-agent reinforcement learning, the exploration-exploitation trade-off is further complicated by the non-stationarity induced by concurrently learning agents. Several studies address this challenge by adopting decentralized, asynchronous, or coordination-aware exploration strategies that reduce interference between agents while preserving adaptability [19, 26, 29]. These approaches allow agents to explore independently or at different time scales, mitigating instability caused by simultaneous exploration. As a result, coordinated behavior can emerge despite dynamically changing interaction patterns and partially observable environments.

3.2.5 Resource-Aware and Real-Time Exploration Control

In real-time and resource-constrained settings, exploration strategies are explicitly limited by computational, energy, or safety constraints. Several works prioritize conservative exploration by restricting exploratory actions to ensure real-time feasibility and system stability [6, 20, 23, 30]. In these scenarios, exploration is carefully modulated to avoid performance degradation or unsafe behavior, particularly in physical systems such as robots and UAVs. This resource-aware regulation of exploration enables online adaptation while respecting strict operational constraints.

3.3 Q3: What are the main applications and reported performance of online RL methods in real-world or simulated autonomous agent scenarios?

The reviewed studies show that online reinforcement learning has been applied to a broad spectrum of autonomous agent scenarios, ranging from robotic manipulation and UAV navigation to multi-agent coordination, communication networks, and energy-aware control. Most evaluations are conducted in simulation environments, while a smaller but growing subset of works demonstrates feasibility under real-time or hardware-related constraints.

Across these application domains, online RL consistently enables agents to cope with unknown dynamics, nonstationary environments, and evolving task requirements more effectively than static or purely offline-trained policies. Actor-critic and model-free adaptive control methods dominate continuous control tasks, whereas continual and lifelong learning approaches emphasize sustained performance over extended deployments. Offline-to-online adaptation further reduces deployment risks by leveraging offline knowledge while preserving online adaptability.

In distributed and multi-agent scenarios, online MARL approaches support scalable coordination and consensus under agent-induced nonstationarity. Complementary paradigms, including uncertainty-aware, knowledge-integrated, and skill-based methods, further improve robustness and learning efficiency in complex or safety-critical settings.

Overall, the surveyed evidence indicates that online policy adaptation is a key

Table 3.1: Main application domains of online reinforcement learning methods (Q3).

Category	References	Application Domain
Online Actor-Critic Based Methods	[14, 15, 9, 22, 5, 6, 27]	Continuous control of robots and UAVs, including robotic manipulators, fixed-wing UAV flight control, formation control, collision avoidance, and autonomous maneuver decision-making
Model-Free Online Adaptive Control	[7, 10, 17, 24]	Distributed control, consensus and tracking in multi-agent systems, adaptive control under unknown and time-varying dynamics, and energy management applications
Continual and Lifelong Online RL	[2, 3, 4, 16, 13]	Long-term robotic control, UAV-assisted IoT networks, navigation in changing environments, and autonomous driving in partially observable scenarios
Offline-to-Online Policy Adaptation	[6, 30, 13, 1]	Simulation-to-real transfer, UAV navigation, visual tracking, and autonomous driving under deployment constraints
Uncertainty-Aware Online Policy Adaptation	[31]	Robust multi-agent decision-making under partial observability and nonstationary dynamics
Knowledge-Integrated Online RL	[21, 16]	Navigation and coordination tasks leveraging prior knowledge or domain constraints
Skill-Conditioned and Hierarchical Online RL	[8, 26]	Hierarchical navigation, curriculum-based learning, and coordinated UAV flocking
Online Multi-Agent RL	[7, 25, 19, 29, 31]	Consensus, formation control, scheduling, routing, and cooperative multi-agent decision-making
Real-Time / Pure Online RL	[23, 20, 6]	Real-time robotic control, vision-based robotics, servo control, and hardware-constrained learning systems

enabler for practical autonomy. Although large-scale real-world validation remains limited in some domains, reported results consistently highlight improved robustness, flexibility, and long-term performance compared to non-adaptive baselines.

Table 3.2: Reported performance of online reinforcement learning methods across application domains (Q3).

Category	**References**	**Reported Performance**
Online Actor-Critic Based Methods	[14, 15, 9, 22]	Stable policy convergence, accurate trajectory tracking, and improved control performance compared to classical or non-adaptive controllers
Model-Free Online Adaptive Control	[7, 10, 17, 24]	Reduced tracking error, convergence under unknown dynamics, and improved adaptability in time-varying environments
Continual and Lifelong Online RL	[2, 3, 4, 16, 13]	Sustained performance over long time horizons, mitigation of catastrophic forgetting, and robustness to nonstationary task sequences
Offline-to-Online Policy Adaptation	[6, 30, 13]	Improved robustness and safety during deployment, better generalization to unseen conditions compared to offline-only policies
Uncertainty-Aware Online Policy Adaptation	[31]	Enhanced robustness and learning stability under uncertainty, with improved average performance in nonstationary environments
Knowledge-Integrated Online RL	[21, 16]	Faster convergence, reduced exploration cost, and improved coordination compared to purely data-driven online RL
Skill-Conditioned and Hierarchical Online RL	[8, 26]	Improved learning efficiency, higher task success rates, and better scalability compared to flat policy representations
Online Multi-Agent RL	[7, 25, 19, 29]	Scalable coordination, robustness to agent-induced nonstationarity, and improved performance over centralized or static baselines
Real-Time / Pure Online RL	[23, 20]	Feasible real-time learning with stable control performance under strict timing and computational constraints

Chapter 4

Conclusion

Online reinforcement learning has emerged as a central paradigm for enabling autonomous agents to operate effectively in dynamic, uncertain, and nonstationary environments. By allowing policies to be adapted during execution, online learning shifts the focus from static optimization to continuous decision-making, which is increasingly required in real-world autonomous systems.
The literature reviewed in this work indicates that the practical relevance of online reinforcement learning extends beyond algorithmic innovation. Across diverse application domains, the ability to learn during deployment supports robustness to model inaccuracies, environmental changes, and evolving task requirements. Rather than relying on a single dominant methodology, current research reflects a trend toward hybrid solutions that combine online learning with structural priors, safety mechanisms, and domain knowledge to achieve reliable behavior under operational constraints.
Despite its potential, online reinforcement learning remains characterized by open challenges. Issues related to safety, scalability, and evaluation in real-world settings continue to limit broader adoption, and the gap between simulation-based validation and long-term deployment persists in many application areas. Addressing these challenges will require closer integration between learning algorithms, control theory, and system-level considerations, as well as the development of standardized benchmarks and evaluation protocols.

In conclusion, online reinforcement learning represents a promising direction for advancing autonomous decision-making in complex environments. Continued progress in this field is likely to play a key role in the development of autonomous systems capable of long-term, adaptive, and reliable operation under real-world conditions.

Bibliography

[1] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang. Tacsl: A library for visuotactile sensor simulation and learning. *IEEE Transactions on Robotics*, 41:2645–2661, 2025. cited By 4.

[2] I. Ganie and S. Jagannathan. Lifelong reinforcement learning tracking control of nonlinear strict-feedback systems using multilayer neural networks with constraints. *Neurocomputing*, 600, 2024. cited By 5.

[3] I. Ganie and S. Jagannathan. Online continual safe reinforcement learning-based optimal control of mobile robot formations. pages 519–524, 2024. cited By 0.

[4] Z. Gong, O. Hashash, Y. Wang, Q. Cui, W. Ni, W. Saad, and K. Sakaguchi. Uav-aided lifelong learning for aoi and energy optimization in nonstationary iot networks. *IEEE Internet of Things Journal*, 11(24):39206–39224, 2024. cited By 11.

[5] Z. Hosny, A. Nassar, A. Aboelyazeed, M. Mohamed, M. Abouheaf, and W. Gueaieb. An online model-free reinforcement learning approach for 6-dof robot manipulators. 2023. cited By 1.

[6] S. Hu, X. Yuan, W. Ni, X. Wang, and A. Jamalipour. Visual-based moving target tracking with solar-powered fixed-wing uav: A new learning-based approach. *IEEE Transactions on Intelligent Transportation Systems*, 25(8):9115–9129, 2024. cited By 19.

[7] X. Ji, X. Zhang, S. Zhu, F. Deng, and B. Zhu. Data-driven adaptive consensus control for heterogeneous nonlinear multi-agent systems using online reinforcement learning. *Neurocomputing*, 596, 2024. cited By 6.

[8] S. Karimi, S. Asadi, and A.H. Payberah. Score: Skill-conditioned online reinforcement learning. volume 20, pages 189–198, 2024. cited By 0.

[9] F. Kong, Z. Zhao, and L. Cheng. Design of adaptive learning control of fixed-wing uav based on actor-critic. volume 2023-July, pages 2231–2236, 2023. cited By 2.

[10] P. Li, W. Zou, J. Guo, and Z. Xiang. Optimal consensus of a class of discrete-time linear multi-agent systems via value iteration with guaranteed admissibility. *Neurocomputing*, 516:1–10, 2023. cited By 20.

[11] Q. Li, B. Li, Y. Rong, Z.-Q. He, and Z. Han. Uav-enabled integrated sensing, communication, and control: A constrained rl approach. *IEEE Internet of Things Journal*, 12(24):53689–53703, 2025. cited By 0.

[12] C. Liu, P. Wu, M. Xu, Y. Yang, and N. Geng. Scalable deep reinforcement learning-based online routing for multi-type service requirements. *IEEE Transactions on Parallel and Distributed Systems*, 34(8):2337–2351, 2023. cited By 29.

[13] A.Q. Md, D. Jaiswal, S. Mohan, N. Innab, R. Sulaiman, M.K. Alaoui, and A. Ahmadian. A novel approach for self-driving car in partially observable environment using life long reinforcement learning. *Sustainable Energy Grids and Networks*, 38, 2024. cited By 4.

[14] Z. Ming, H. Zhang, J. Zhang, and X. Xie. A novel actor–critic–identifier architecture for nonlinear multiagent systems with gradient descent method. *Automatica*, 155, 2023. cited By 34.

[15] P. Osinenko, G. Yaremenko, G. Malaniya, and A. Bolychev. An actor-critic framework for online control with environment stability guarantee. *IEEE Access*, 11:89188–89204, 2023. cited By 1.

[16] P. Qin, J. Zhao, Z. Mei, H. Yang, and T. Zhao. Knowledge guided continual reinforcement learning for navigation. *IEEE Internet of Things Journal*, 2025. cited By 0.

[17] K. Rao, H. Yan, Q. Liu, Q. Dang, and K. Shi. Optimal tracking control of second-order multiagent systems with input delay via data-driven forward reward q-learning framework. *IEEE Transactions on Systems Man and Cybernetics Systems*, 55(3):1858–1869, 2025. cited By 1.

[18] J. Ren, Y. Lan, X. Xu, Y. Zhang, Q. Fang, and Y. Zeng. Deep reinforcement learning using least-squares truncated temporal-difference. *Caai Transactions on Intelligence Technology*, 9(2):425–439, 2024. cited By 4.

[19] A. Srinivasan, J. Zhang, and O. Tirkkonen. Asynchronous multi-agent reinforcement learning for scheduling in subnetworks. 2025. cited By 0.

[20] L. Stiemer, M.M. Groves-Raines, L. Wood, A. Mohamed, and T. Wiley. Online deep reinforcement learning of servo control for a small-scale bio-inspired wing. *Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 15443 LNAI:65–76, 2025. cited By 1.

[21] Hainan Tang, Hongjie Tang, Juntao Liu, Ziyun Rao, Yunshu Zhang, and Xunhao Luo. A configuration of multi-agent reinforcement learning integrating prior knowledge. 2024. cited By 0.

[22] C.-C. Tsai, H.-Y. Chen, S.-C. Chen, F.-C. Tai, and G.-M. Chen. Adaptive reinforcement learning formation control using orfbls for omnidirectional mobile multi-robots. *International Journal of Fuzzy Systems*, 25(5):1756–1769, 2023. cited By 8.

[23] Y. Wang, G. Vasan, and A.R. Mahmood. Real-time reinforcement learning for vision-based robotics utilizing local and remote computers. volume 2023-May, pages 9435–9441, 2023. cited By 9.

[24] Z. Wang, C. Chen, and D. Dong. Instance weighted incremental evolution strategies for reinforcement learning in dynamic environments. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9742–9756, 2023. cited By 12.

[25] S. Xuan, H. Liang, S. Huang, T. Li, and J. Sun. Distributed optimal consensus problem of input constrained nonlinear discrete-time mass: A mode-free reinforcement learning approach. *IEEE Transactions on Cybernetics*, 55(6):2910–2923, 2025. cited By 3.

[26] C. Yan, C. Wang, X. Xiang, K.H. Low, X. Wang, X. Xu, and L. Shen. Collision-avoiding flocking with multiple fixed-wing uavs in obstacle-cluttered environments: A task-specific curriculum- based madrl approach. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10894–10908, 2024. cited By 62.

[27] S. Yang, H. Li, Y. Hong, C. Chen, and G. Liu. Autonomous maneuver decision-making for close air combat based on lstm-ddpg algorithm. pages 1952–1959, 2025. cited By 0.

[28] Q. Yuan. Residential demand response online optimization based on multi-agent deep reinforcement learning. *Electric Power Systems Research*, 237, 2024. cited By 5.

[29] Z. Zhang, B. Zhang, G. Zhou, D. Li, Z. Xu, and G. Fan. Decentralized extension for centralized multi-agent reinforcement learning via online distillation. *Lecture Notes in Computer Science*, 15288 LNCS:211–225, 2025. cited By 0.

[30] S. Zhao, F. Zhou, Q. Wu, and F. Shen. Energy-effcient uav coverage aware navigation under continuous dynamic constraints: An offline-online radio map-enhanced drl method. *IEEE Transactions on Communications*, 2025. cited By 0.

[31] X. Zhao, J. Liu, F. Wu, X. Zhang, and G. Wang. Uncertainty modified policy for multi-agent reinforcement learning. *Applied Intelligence*, 54(22):12020–12034, 2024. cited By 1.