



ONLINE REINFORCEMENT LEARNING FOR DECISION-MAKING AND EXPLICIT POLICY LEARNING IN AUTONOMOUS AGENTS

A Systematic Literature Review

Pablo Sebastian Vargas Grateron
Course of Intelligent Systems Engineering

February 2026



WHY ONLINE REINFORCEMENT LEARNING?

Many RL approaches rely on **offline or pre-trained policies**.

Real autonomous systems operate in:

- dynamic environments
- non-stationary conditions
- interaction with other agents

Policies must **adapt during execution**.

OBJECTIVES OF THE SLR

- Identify truly online reinforcement learning approaches that **update policies during execution**, distinguishing them from offline or static methods.
- Analyze how autonomous agents handle environmental **non-stationarity, multi-agent interactions**, and **complex dynamics** in online decision-making.
- Map the main application domains (e.g., robotics, UAVs, smart grids, autonomous vehicles).



RESEARCH QUESTIONS

RQ1: Which online reinforcement learning (RL) methods have been proposed to enable dynamic policy adaptation in autonomous agents during execution?

RQ2: Which strategies are adopted to balance exploration and exploitation in dynamic and non-stationary environments for online learning?

RQ3: What are the main applications and reported performance of online RL methods in real-world or simulated autonomous agent scenarios?



METHODOLOGY

Database: Scopus (broad coverage in Computer Science & Engineering; support for Boolean and proximity operators)

Search scope: TITLE-ABS-KEY fields, limited to Computer Science and Engineering

Focus: Online reinforcement learning methods with continuous or incremental policy adaptation

Exclusions: Offline/batch RL, imitation-based methods, off-policy evaluation

Study Selection Process

Multi-stage selection to identify relevant primary studies and reduce bias.

Stage 1 – Screening: title, abstract, and metadata.

Stage 2 – Full-text review: methodological relevance and scope alignment.

Stage 3 – Final selection: studies included in the qualitative synthesis.

Inclusion criteria:

- **Online, policy-based RL** with updates during execution
- Dynamic / **non-stationary environments**
- **Autonomous and robotic systems**
- Peer-reviewed, English publications (2023–2026), pre-2023 only if >50 citations.

Exclusion criteria:

- Offline or batch RL
- Imitation-based or non-policy methods
- No explicit online policy adaptation
- Non-peer-reviewed or methodologically insufficient studies

STUDY SELECTION



RESULTS



RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?



+

- 1. Online Actor-Critic Based Methods
- 2. Model-Free Online Adaptive Control via Reinforcement Learning
- 3. Continual and Lifelong Online Reinforcement Learning
- 4. Offline-to-Online Policy Adaptation
- 5. Uncertainty-Aware Online Policy Adaptation
- 6. Knowledge-Integrated Online Reinforcement Learning
- 7. Skill-Conditioned and Hierarchical Online Reinforcement Learning
- 8. Online Multi-Agent Reinforcement Learning (MARL)
- 9. Real-Time / Pure Online Reinforcement Learning

RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

1. Online Actor–Critic Based Methods

These methods continuously update both the **actor (policy)** and the **critic (value estimator)** during execution, allowing agents to refine their behavior step by step as new data arrives.

Because **the actor directly adjusts its actions based on real-time feedback from the critic**, these approaches provide stable and responsive adaptation in continuous-control settings.

Variants include deep actor–critic networks, safety-aware formulations using stability constraints, and decentralized versions for multi-agent scenarios.

RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

2. Model-Free Online Adaptive Control via Reinforcement Learning

Model-free adaptive RL methods enable online policy improvement **without requiring explicit mathematical models of system dynamics**.

Agents update their control parameters directly **from observed transitions** using value-based or policy-gradient techniques, making them **well suited to environments where modeling is difficult or dynamics are time-varying**.

Their strength lies in providing robust adaptation to real-world uncertainties while avoiding the complexity of model identification.

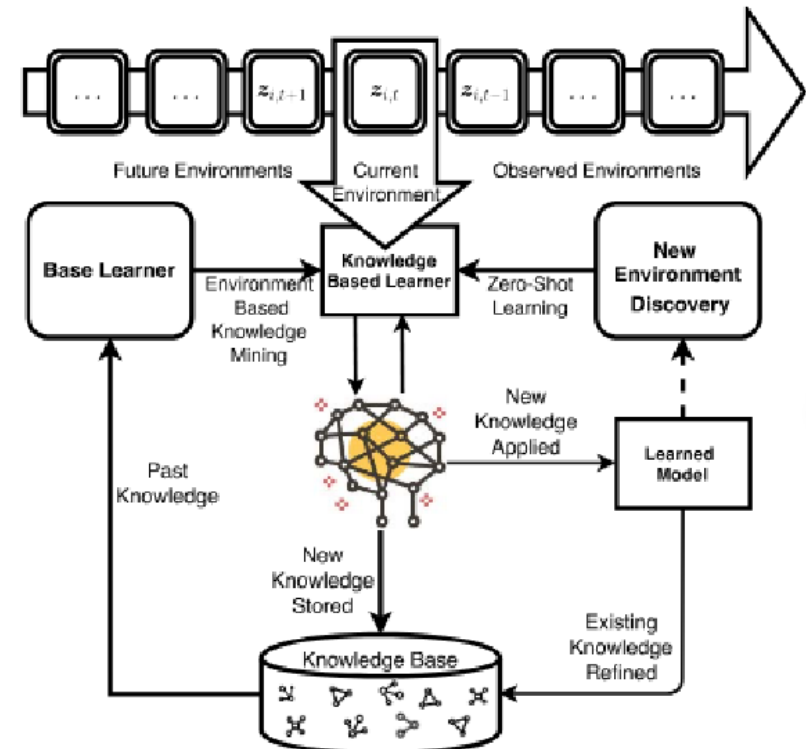
RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

3. Continual and Lifelong Online Reinforcement Learning

Continual and lifelong RL approaches focus on **enabling agents to learn over extended periods while retaining previously acquired skills** and adapting to new tasks or changes.

These methods integrate mechanisms such as knowledge preservation, selective regularization, and incremental updates to avoid catastrophic forgetting.

As a result, policies remain stable yet adaptable, evolving smoothly as the environment or mission requirements shift.



RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

4. Offline-to-Online Policy Adaptation

These methods **combine offline training with online refinement** to ensure safe but flexible behavior during deployment.

A policy is first **trained using simulation or historical data**, then **adjusted online to account for real-world discrepancies** such as noise, unmodeled dynamics, or nonstationarity.

This hybrid approach reduces exploration risks while still enabling dynamic adaptation in unpredictable environments.

RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

5. Uncertainty-Aware Online Policy Adaptation

Uncertainty-aware approaches **incorporate explicit estimates of uncertainty into the policy update process**, allowing agents to modulate their decisions based on confidence in their current value or policy predictions.

When uncertainty is high, exploration is encouraged; when it decreases, exploitation is favored.

This leads to more robust behavior in partially observable or multi-agent settings where unpredictability is inherent.

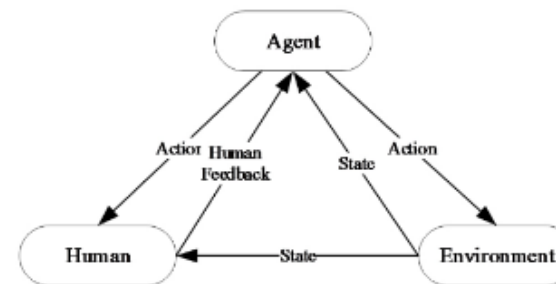
RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

6. Knowledge-Integrated Online Reinforcement Learning

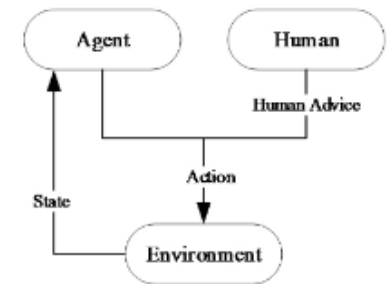
In knowledge-integrated methods, **prior knowledge** such as constraints, heuristics, or task-specific structure is **used to guide and shape online policy adaptation**.

This additional information helps reduce unnecessary exploration, improve stability, and accelerate convergence, especially in safety-critical or complex environments.

The agent still learns online, **but its updates are constrained or informed** by external domain insights.



(a) Interactive Reinforcement Learning



(b) Advice-added Reinforcement Learning

RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

7. Skill-Conditioned and Hierarchical Online Reinforcement Learning

These approaches introduce **structured decision-making** by decomposing tasks into reusable skills or organizing policies hierarchically.

A high-level policy selects among skills, while low-level controllers adapt execution online.

This structure improves scalability and modularity, enabling rapid adaptation to new tasks or conditions without relearning all behaviors from scratch.

RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

8. Online Multi-Agent Reinforcement Learning (MARL)

Online MARL methods **allow multiple agents to learn concurrently while interacting in shared environments**, addressing challenges like non-stationarity caused by other learning agents.

Policies are often decentralized or asynchronously updated to reduce interference and improve stability.

This enables scalable and adaptive coordination for tasks such as formation control, scheduling, or distributed decision-making.

RQ1: WHICH ONLINE REINFORCEMENT LEARNING (RL) METHODS HAVE BEEN PROPOSED TO ENABLE DYNAMIC POLICY ADAPTATION IN AUTONOMOUS AGENTS DURING EXECUTION?

9. Real-Time / Pure Online Reinforcement Learning

Real-time or pure online RL methods **learn almost entirely during execution**, with minimal reliance on offline data, and **are designed to operate under strict timing, hardware, or energy constraints**.

They emphasize fast, lightweight updates to ensure safe and stable behavior during continuous real-world operation.

These approaches demonstrate that genuine online learning is feasible even in tightly constrained robotic and UAV applications.

RO2: WHICH STRATEGIES ARE ADOPTED TO BALANCE EXPLORATION AND EXPLOITATION IN DYNAMIC AND NON-STATIONARY ENVIRONMENTS FOR ONLINE LEARNING?



+

- 1. Adaptive Exploration Scheduling
- 2. Uncertainty-Driven Exploration
- 3. Continual and Lifelong Exploration Strategies
- 4. Exploration Strategies in Multi-Agent and Non-Stationary Settings
- 5. Resource-Aware and Real-Time Exploration Control



RQ2: WHICH STRATEGIES ARE ADOPTED TO BALANCE EXPLORATION AND EXPLOITATION IN DYNAMIC AND NON-STATIONARY ENVIRONMENTS FOR ONLINE LEARNING?

1. Adaptive Exploration Scheduling:

Exploration is adjusted online based on **performance trends** or **environmental changes**.

→ When performance drops or the environment shifts → **more exploration**.

→ When behavior stabilizes → **more exploitation**.

2. Uncertainty-Driven Exploration:

Exploration is regulated using **uncertainty estimates** of the policy, value function, or environment.

→ **Reduces overconfident exploitation** and improves robustness in dynamic or multi-agent settings.



RQ2: WHICH STRATEGIES ARE ADOPTED TO BALANCE EXPLORATION AND EXPLOITATION IN DYNAMIC AND NON-STATIONARY ENVIRONMENTS FOR ONLINE LEARNING?

3. Continual and Lifelong Exploration Strategies

Exploration is never fully annealed; it is reactivated when new tasks or changes appear.

→ Prevents premature convergence and mitigates catastrophic forgetting.

4. Multi-Agent / Non-Stationary Exploration

Agents **explore asynchronously or independently** to reduce interference and instability.

→ Supports coordination despite non-stationarity induced by other learning agents.



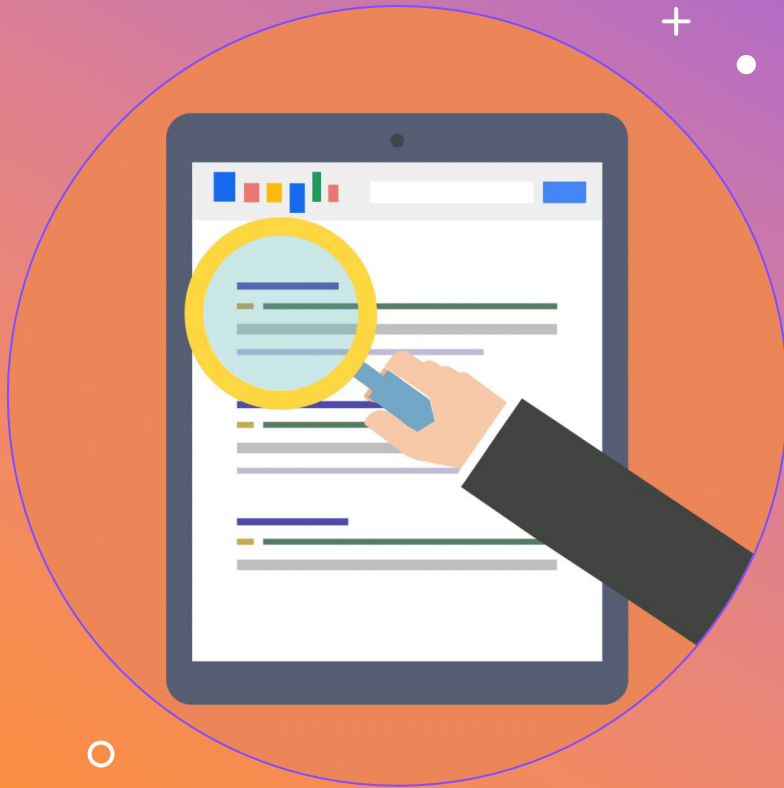
**RQ2: WHICH STRATEGIES ARE ADOPTED TO BALANCE EXPLORATION AND EXPLOITATION
IN DYNAMIC AND NON-STATIONARY ENVIRONMENTS FOR ONLINE LEARNING?**

5. Resource-Aware and Real-Time Exploration Control

Exploration is **explicitly constrained by computational, energy, and safety limits** to ensure real-time feasibility and stable online adaptation.

→ Keeps exploration conservative to guarantee real-time feasibility and avoid unsafe behavior.

**RQ3: WHAT ARE THE MAIN APPLICATIONS AND REPORTED PERFORMANCE OF
ONLINE RL METHODS IN REAL-WORLD OR SIMULATED AUTONOMOUS AGENT SCENARIOS?**



RQ3: WHAT ARE THE MAIN APPLICATIONS AND REPORTED PERFORMANCE OF

ONLINE RL METHODS IN REAL-WORLD OR SIMULATED AUTONOMOUS AGENT SCENARIOS?

Category	Application Domain	Reported Performance
Online Actor-Critic Based Methods	• Continuous robot control • UAV flight control • Formation control & collision avoidance • Autonomous maneuvering	• Stable policy convergence • Accurate trajectory tracking • Better control performance than non-adaptive baselines
Model-Free Online Adaptive Control	• Distributed multi-agent control • Consensus & tracking • Adaptive control under unknown/time-varying dynamics • Energy-management tasks	• Reduced tracking error • Convergence under unknown dynamics • Improved adaptability in time-varying environments
Continual and Lifelong Online RL	• Long-term robotic control • UAV-assisted IoT systems • Navigation in changing environments • Autonomous driving in partially observable settings	• Sustained long-term performance • Mitigation of catastrophic forgetting • Robustness to nonstationary task sequences

RQ3: WHAT ARE THE MAIN APPLICATIONS AND REPORTED PERFORMANCE OF

ONLINE RL METHODS IN REAL-WORLD OR SIMULATED AUTONOMOUS AGENT SCENARIOS?

Category	Application Domain	Reported Performance
Offline-to-Online Policy Adaptation	• Sim-to-Real transfer • UAV navigation • Visual tracking • Autonomous driving under deployment constraints	• More robust deployment • Better generalization to unseen conditions • Reduced risk vs. offline-only policies
Uncertainty-Aware Online Policy Adaptation	• Robust multi-agent decision-making • Partial observability • Nonstationary dynamics	• Higher robustness under uncertainty • Improved learning stability • Better average performance in nonstationary environments
Knowledge-Integrated Online RL	• Navigation with prior knowledge • Multi-agent coordination with domain constraints	• Faster convergence • Lower exploration cost • Improved coordination using prior knowledge

RQ3: WHAT ARE THE MAIN APPLICATIONS AND REPORTED PERFORMANCE OF

ONLINE RL METHODS IN REAL-WORLD OR SIMULATED AUTONOMOUS AGENT SCENARIOS?

Category	Application Domain	Reported Performance
Skill-Conditioned & Hierarchical Online RL	• Hierarchical navigation • Curriculum-based learning • Coordinated UAV flocking	• Higher learning efficiency • Increased task success rates • Better scalability than flat policies
Online Multi-Agent RL (MARL)	• Consensus • Formation control • Scheduling & routing • Cooperative multi-agent decision-making	• Scalable coordination • Robustness to agent-induced nonstationarity • Outperforms centralized/static baselines
Real-Time / Pure Online RL	• Real-time robotic control • Vision-based robotics • Servo control • Hardware-constrained learning	• Feasible real-time learning • Stable control under strict timing constraints • Consistent performance on hardware-limited systems

CONCLUSION

Online reinforcement learning enables autonomous agents to adapt continuously to dynamic and uncertain environments, outperforming static or offline policies. Its growing impact across robotics, UAVs, and multi-agent systems highlights the value of hybrid, safety-aware, and knowledge-integrated approaches.

Despite challenges in safety, scalability, and real-world evaluation, online RL remains a key pathway toward autonomous systems capable of long-term, robust, and reliable operation.





x

•

o

END