In: 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC),
05-08 October 2025, Vienna, Austria, pp. 5655-5660
DOI: 10.1109/SMC58881.2025.11342671

Learning What Matters Now: A Dual-Critic Context-Aware RL Framework for Priority-Driven Information Gain

Dimitris Panagopoulos[1], Adolfo Perrusquía[1] and Weisi Guo[1]

***Abstract*— Autonomous systems operating in high-stakes search-and-rescue (SAR) missions must continuously gather mission-critical information while flexibly adapting to shifting operational priorities. We propose CA-MIQ (Context-Aware Max-Information Q-learning), a lightweight dual-critic reinforcement learning (RL) framework that dynamically adjusts its exploration strategy whenever mission priorities change. CA-MIQ pairs a standard extrinsic critic for task reward with an intrinsic critic that fuses state-novelty, information-location awareness, and real-time priority alignment. A built-in shift detector triggers transient exploration boosts and selective critic resets, allowing the agent to re-focus after a priority revision. In a simulated SAR grid-world, where experiments specifically test adaptation to changes in the priority order of information types the agent is expected to focus on, CA-MIQ achieves nearly four times higher mission-success rates than baselines after a single priority shift and more than three times better performance in multiple-shift scenarios, achieving 100% recovery while baseline methods fail to adapt. These results highlight CA-MIQ's effectiveness in any discrete environment with piecewise-stationary information-value distributions.**

***Index Terms*— Information gain, intrinsic motivation, priority shift, reinforcement learning**

I. Introduction

Search and rescue (SAR) operations in disaster scenarios present extraordinary challenges for autonomous agents. These environments are inherently dynamic, with conditions that evolve rapidly as new information becomes available. Traditional reinforcement learning (RL) methods for robotic exploration, whether ε-greedy, Boltzmann, or even maximum-entropy variants, often assume a stationary environment with fixed priorities, failing to account for the fluid nature of SAR missions where the relative importance of information types can shift abruptly as the situation evolves [1], [2]. For instance, the discovery of a gas leak may suddenly elevate "hazard assessment B" to higher priority than "victim status A." Similarly, the identification of an unstable structure might necessitate immediate exploration of alternative access routes. Existing approaches to information-directed exploration, while effective in stationary environments, lack mechanisms for adapting to these priority shifts [3], [4]. This results in agents that continue sampling from distributions no longer aligned with operational needs leading to wasted time, unnecessary risk exposure, and suboptimal triage outcomes.

Recent advances have begun addressing these challenges through different approaches. Panagopoulos et al. [5] combine a hierarchical decision-making architecture with RL and large language model (LLM) that formalises information gathering task within a structured Information Space. While this formulation enables agents to collect information in a systematic manner, it assumes a static ordering of priorities throughout the mission. This limitation significantly restricts the practical utility of autonomous agents in real-world disaster scenarios where responders must continuously reassess and reprioritise objectives based on emerging information. In a complementary direction, the authors in [3] demonstrate that exploration strategies maximising information gain significantly improve sample-efficiency in continuous control tasks by directing agents toward transitions with highest uncertainty. However, their framework is formulated for high-dimensional function-approximation settings and does not address mission-level priority shifts in the discrete, tabular domains that characterise high-level decision-making [6].

These limitations reveal a critical gap. Systems guided by human knowledge can incorporate and exploit stakeholder bias but cannot effectively re-explore when information is revised, leading to rigid behavior that fails to adapt to changing conditions [7], [8]. While information-gain driven methods can trigger purposeful exploration, they lack mechanisms for aligning that exploration with dynamic, task-level priorities, resulting in agents that explore efficiently but not necessarily in directions aligned with current mission objectives. Bridging this gap requires an exploration policy that (1) recognises when its current ordering of information targets becomes obsolete, and (2) focuses exploration effort toward prioritised transitions without discarding still-relevant knowledge.

To address these limitations, we introduce CA-MIQ (Context-Aware Max-Information Q-learning), a framework unifying information-gain maximisation with priority-aware adaptation in discrete state-action environments. Our approach treats priority ordering as a latent context variable that can shift dynamically, posing the key question: *How do you keep exploring the right facts when mission priorities keep shuffling?* As shown in Figure 1, CA-MIQ builds on [3] and incorporates three main innovations. First, we extend the Information Space formulation [5] by modeling priority ordering explicitly as a latent context, enabling the agent to detect and adapt to shifts via performance statistics or operator input. Second, CA-MIQ maintains two tabular critics: an extrinsic critic for task rewards and an intrinsic critic for information-gain, balancing state novelty, information-location awareness, and priority alignment. Unlike existing methods (e.g., MaxInfoRL [3]) that reduce global uncer-

[1]Faculty of Engineering and Applied Science, Cranfield University, Cranfield MK43 0AL, UK, {d.panagopoulos, adolfo.perrusquia-guzman, weisi.guo}@cranfield.ac.uk. This work is funded by EPSRC iCASE with Thales UK (EP/X52475X/1)

Published by IEEE. This is the Author Accepted Manuscript issued with: Creative Commons Attribution License (CC:BY 4.0).
The final published version (version of record) is available online at DOI:10.1109/SMC58881.2025.11342671. Please refer to any applicable publisher terms of use.

tainty, we focus specifically on high-level, mission-critical information in shifting priority contexts. Third, upon priority shifts, our framework temporarily boosts exploration through a transient ε-increase while selectively resetting affected critic entries.

Experimental results show CA-MIQ significantly outperforms traditional tabular methods, achieving optimal post-shift performance nearly four and three times faster in single and multiple shift scenarios, respectively. By bridging static information structures and dynamic operational realities, our work represents the first information-gain-driven exploration strategy addressing piecewise-stationary priority shifts in SAR environments.

II. Related Work

Traditional exploration methods in RL include ε-greedy [9], Boltzmann exploration [10], and Upper Confidence Bound (UCB) [11], despite their theoretical guarantees, they struggle to adapt to scenarios with shifting priorities. Similarly, Thompson sampling methods [11] also assume fixed underlying reward distributions, making them unsuitable when priorities shift dynamically.

Intrinsic motivation has emerged as a powerful paradigm for exploration, leveraging curiosity [12], novelty [13], and surprise [14] as intrinsic rewards. Notably, Burda et al. [15] demonstrated curiosity-driven methods could effectively address sparse-reward scenarios by promoting exploration of unfamiliar state-action spaces. However, intrinsic motivation approaches typically optimise for epistemic uncertainty globally, lacking explicit mechanisms to align exploration with dynamically changing operational priorities. As a result, agents may continue exploring states or actions irrelevant to mission-critical objectives despite strong intrinsic signals, causing inefficiencies or mission failures in operational contexts.

Adapting to changing contexts remains challenging in RL. Meta-reinforcement learning [16] aims to rapidly adapt to new tasks, while contextual bandits [17] explicitly model context variables to enhance decision-making. Hidden-parameter Markov Decision Processes (MDPs) [18] formalise environments with shifting underlying parameters but typically assume contexts are fully observable or can be inferred solely from environmental dynamics. Recent works have explicitly modeled context switches [19] and developed context-aware exploration methods that adapt to dynamic reward structures, primarily in continuous domains using deep RL [20]. However, such methods rarely address discrete, partially observable contexts typical of high-level decision-making in SAR operations, where shifts often occur explicitly via human input or implicitly through subtle performance degradation.

Information-theoretic exploration approaches provide principled mechanisms for balancing information gain with exploitation. Information-Directed Sampling (IDS) [21] balances expected regret against information gain about the optimal action. Maximum entropy RL [22] encourages exploration by maximising policy entropy alongside expected return. While effective in stationary settings, these methods are not inherently designed to respond to abrupt shifts in operational priorities.

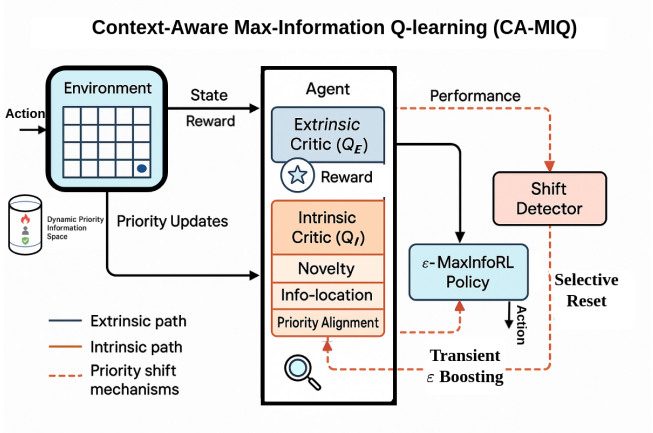


Fig. 1: **CA-MIQ (Context-Aware Max-Information Q-learning) architecture.** The environment supplies the current state to an agent equipped with two critics: an *extrinsic critic* Q^E that evaluates task reward-seeking behaviour (black path) and an *intrinsic critic* Q^I that aggregates novelty, information-location and priority-alignment signals (orange path). An ε-MaxInfoRL policy gate arbitrates between the critics to choose each action. A *shift detector* monitors performance; when a distributional change is detected it triggers a dashed orange *priority-shift loop* that transiently boosts exploration (ε-boosting) and *resets* selectively the critics. Solid arrows denote data flow, whereas dashed arrows denote adaptive control signals.

To address these limitations, CA-MIQ explicitly integrates information-gain maximisation with priority-driven intrinsic rewards and explicit context-shift detection. Unlike purely curiosity-driven or intrinsically motivated methods, CA-MIQ strategically aligns exploration behaviors with explicitly shifting mission priorities, ensuring targeted information gathering aligned with current operational objectives. Consequently, CA-MIQ quickly adapts exploration behaviors in discrete, piecewise-stationary environments, offering practical advantages over traditional, curiosity-driven, and purely information-theoretic methods in SAR contexts.

III. Methodology

A. Problem Formulation

We consider a tabular RL setting in a SAR scenario, modeled as an MDP $\mathcal{M} = (S, A, P, R, \gamma)$, where S is the state space, A is the action space, $P : S \times A \times S \to [0, 1]$ is the transition function, $R : S \times A \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. Unlike standard MDPs, we explicitly model the collection of different types of information as a critical component of the agent's task. Similalry to [5], this is formalised through an Information Space $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$, where each i_k represents a specific type of information (e.g., victim location, hazard assessment, safe zone) that the agent can collect. Crucially, each information

type has an associated priority $p_k \in \mathbb{R}^+$ that determines its relative importance. The collection priority ordering is initially defined as a mapping $\mathcal{O}_t : \mathcal{I} \to \{1, 2, \ldots, |\mathcal{I}|\}$, where $\mathcal{O}_t(i_k)$ indicates the position of information type i_k in the collection sequence. Each item $i_k \in \mathcal{I}$ resides at a location and is labelled with the current order $\mathcal{O}_t(i_k)$. This ordering may change during operation, formally when $\mathcal{O}_t \neq \mathcal{O}_{t-1}$, reflecting the evolving nature of SAR missions.

B. Dynamic-Priority Information Space

Our approach introduces a latent context variable modeling priority ordering, capable of shifting during operation. The information collection task is defined by $(\mathcal{I}, \mathcal{O}, \mathcal{C})$: information types $\mathcal{I}$, current priority ordering $\mathcal{O}$, and feasible priority contexts $\mathcal{C}$. Context shifts ($\Delta\pi$) occur due to operator input or environmental changes at arbitrary time steps. The MDP is conditionally stationary given a fixed context, exhibiting piecewise-stationary dynamics overall. The agent aims to maximise total reward (mission success) and minimise adaptation time after context shifts.

C. Dual-Critic Intrinsic-Extrinsic Architecture

To achieve adaptive information-directed exploration, we borrow the insight of [3] so that the agent keeps two tabular critics: ***Extrinsic critic*** $Q^E(s, a)$ that estimates the expected return from task rewards; and ***Intrinsic critic*** $Q^I(s, a)$ that estimates information gain about the MDP, favoring actions that reduce dynamic, priority-driven uncertainty. The full update equations for the tabular critics are as follows:

$$Q^E(s_t, a_t) \leftarrow Q^E(s_t, a_t) + \alpha[r_t + \gamma \max_{a'} Q^E(s_{t+1}, a') - Q^E(s_t, a_t)] \tag{1}$$

$$Q^I(s_t, a_t) \leftarrow Q^I(s_t, a_t) + \alpha[r_{int}(s_t, a_t) + \gamma \max_{a'} Q^I(s_{t+1}, a') - Q^I(s_t, a_t)] \tag{2}$$

Where $r_{int}(s_t, a_t)$ is the intrinsic reward, computed as a weighted combination of three distinct components:

- ***Novelty-based exploration***: Incentivises visiting rarely tried state–action pairs, preventing premature exploitation and seeking knowledge of alternative routes.

$$r_{\text{novelty}}(s, a) = \frac{\beta_1}{\sqrt{N(s, a)}} \tag{3}$$

  where $N(s, a)$ is the visit count for the state-action pair.
- ***Information location awareness***: Steers exploration toward regions where mission-critical information items are placed, so the agent does not waste effort in redundant and costly exploration.

$$r_{\text{info}}(s, a) = \begin{cases} \frac{\beta_2}{\sqrt{N_{\text{info}}(s)}}, & \text{if } s \text{ is an info location} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

  where $N_{\text{info}}(s)$ tracks visits specifically to information locations.
- ***Priority alignment***: Acts like a compass: it boosts actions that respect the live priority sequence and discourages picking up lower priority facts too early.

$$r_{\text{alignment}}(s, a) = \begin{cases} \beta_3, & \text{if ordered collection} \\ -\beta_4, & \text{if out-of-order collection} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}$ are tunable parameters. The agent's action selection mirrors [3] and is performed using an ε-MaxInfoRL policy:

$$a_t = \begin{cases} \arg\max_a Q^I(s_t, a) & \text{with probability } \varepsilon_t \\ \arg\max_a Q^E(s_t, a) & \text{else,} \end{cases} \tag{6}$$

Unlike traditional exploration rules, our two-critic strategy exploits with Q^E and explores with Q^I, while introducing a priority-weighted intrinsic reward for mission-level alignment in discrete SAR tasks. Since we maintain the standard Bellman update for the Q^E, convergence to the optimal Q-values is guaranteed within each stationary segment, provided sufficient exploration. The Q^I serves primarily as an exploration guide and does not affect the asymptotic convergence properties of the learned policy.

Algorithm 1 Context-Aware Max-Information Q-learning (CA-MIQ)

Require: Learning rate α, discount factor γ, initial exploration ε_0, $\Delta\pi$
Initialize $Q^E(s, a)$ and $Q^I(s, a)$ arbitrarily for all s, a
Initialize priority ordering $\mathcal{O}_0$
for each episode **do**
  Initialize state s
  for each step t **do**
    Select a using Eq. (6))
    Take a, observe r, s'
    Update $Q^E(s, a)$ using Eq. (1)
    Compute r_{int} using Eqs. (3)-(5)
    Update $Q^I(s, a)$ using Eq. (2)
    $s \leftarrow s'$
    if $\Delta\pi$ **then**
      Update $\mathcal{O}$ to new ordering
      Apply selective reset using Eq. (9)
      Boost ε using Eq. (7)
    end if
  end for
  Decay ε using Eq. (8)
end for

D. Priority-Shift Adaptation Mechanism

1) Shift Detection: The agent monitors its collection success rate and maintains historical performance metrics. A context shift is detected either through explicit operator input or when the agent observes a significant drop in collection success alignment.

TABLE I: Performance Comparison (Single Priority Shift)

Metric	Standard Baseline	Boosted Baseline	CA-MIQ (Ours)
Mission Success (%)	18.5	18.2	**65.9**
Info Collection (%)	20.8	20.4	**69.2**
Recovery Success (%)	0.0	0.0	**100.0**
Recovery Time (Ep.)	N/A	N/A	**685**
Reward (per Ep.)	4.2	3.9	**63.9**

2) Transient Epsilon Boosting: Upon detecting a shift, the agent temporarily increases the exploration parameter ε by a factor of λ_{boost}:

$$\varepsilon_t \leftarrow \min(\varepsilon_{\text{MAX}}, \varepsilon_t \cdot \lambda_{\text{boost}}) \tag{7}$$

This adjustment increases the likelihood of selecting the exploratory branch of the ε-MaxInfoRL policy, which relies on the intrinsic critic Q^I. As a result, the agent is more likely to pursue actions driven by information gain to adapt to new priorities. The boosted exploration decays gradually over D_{boost} episodes:

$$\varepsilon_{t+k} = \varepsilon_t \cdot \exp(-k/D_{\text{boost}}), \quad k \in \{1, ..., D_{\text{boost}}\} \tag{8}$$

3) Selective Critic Reset: Rather than discarding all learned values, we apply a partial reset to both critics for all states when paired with information collection actions

$$Q^\bullet(s,a) \leftarrow \lambda \cdot Q^\bullet(s,a), \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}_{\text{collection}} \tag{9}$$

where $\lambda \in \mathbb{R}$ is the reset factor and $\mathcal{A}_{\text{collection}}$ is specifically the set of information collection actions. This approach preserves previously learned knowledge while encouraging re-exploration of collection strategies throughout the entire state space after priority reordering. The pseudocode algorithm is shown in Algorithm 1.

TABLE II: Performance Comparison (Multiple Priority Shifts)

Metric	Standard Baseline	Boosted Baseline	CA-MIQ (Ours)
Mission Success (%)	12.3	13.1	**50.2**
Info Collection (%)	14.5	15.3	**52.9**
Recovery Success (%)	0.0	0.0	**100.0**
Recovery Time (Ep.)	N/A	N/A	**766**
Reward (per Ep.)	-6.1	-5.1	**42.7**

IV. Experiments

A. Environment

To evaluate our approach, we implemented a custom OpenAI Gym gridworld environment , similar to [5], designed to reflect realistic SAR mission dynamics with changing information priorities. The environment features a 4×4 grid where an autonomous agent must collect mission-critical information types (denoted as `X`, `Y`, and `Z`) in a specific priority order before reaching a designated target location to complete the mission. The state and action spaces follow the same structure as in [5]. No hierarchical agents are tested in this setting. A mission is considered successful when the agent collects all required information in the correct order, reaches the target, and saves the victim. The environment assigns rewards for movement, falling into ditches, exceeding action limits at information locations, successful information collection, and completing the mission. Additionally, it provides a method to dynamically swap the priority order of the information types at pre-defined episodes, requiring the agent to detect and adapt its policy without resetting the learned Q-values. The code is available at `https://github.com/dimipan/DynamicPriorities`.

TABLE III: Ablation Study Results (Multiple Priority Shifts)

Configuration	Adapt. Time (Ep.)	Mission Success	Info Collection
Full CA-MIQ	**766**	**50.2%**	**52.9%**
w/o Priority Alignment + Awareness	808.7 (-5.6%) ↓	47.3% (-5.8%) ↓	50.3% (-4.9%) ↓
w/o State Novelty	851.2 (-11.1%) ↓	34.4% (-31.5%) ↓	43.8% (-17.2%) ↓
w/o Exploration Boost	780 (-1.9%) ↓	49.2% (-2.0%) ↓	52.2% (-1.4%) ↓
w/o Selective Reset	777.4 (-1.5%) ↓	49.3% (-1.8%) ↓	52.2% (-1.4%) ↓
Intrinsic Reset Only	770 (-0.5%) ↓	49.5% (-1.4%) ↓	52.7% (-1.3%) ↓
Extrinsic Reset Only	770.3 (-0.6%) ↓	49.6% (-1.2%) ↓	52.7% (-0.5%) ↓

Note: Values in parentheses show the percent change relative to Full CA-MIQ; arrows indicate whether that change is an increase (↑) or a decrease (↓) in performance.

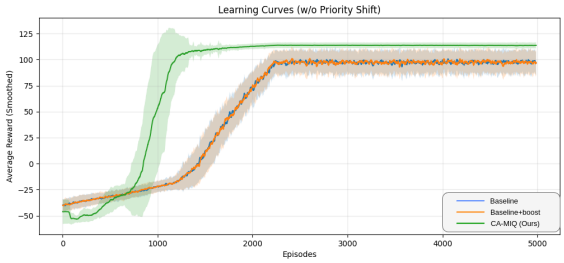
B. Baselines

We use tabular agents because the task is discrete and small providing transparency in analysing how priority shifts affect value estimation and exploration behavior without the added complexity of function approximation. To ensure fairness, we compare CA-MIQ to (i) standard Q-learning, and (ii) Q-learning + ε-Boost, which uses the same boost schedule as CA-MIQ ($D_{\text{boost}} = 50$ episodes with factor $\lambda_{\text{boost}} = 2$) but omits Q^I and resets, isolating the effect of novel and priority-aligned intrinsic exploration.

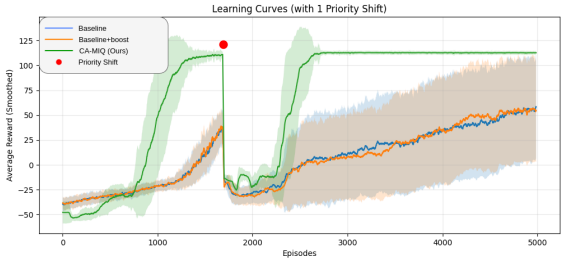
C. Experimental Scenarios

We evaluated our approach in three independent scenarios designed to test adaptation and robustness to dynamic mission demands. First, standard information collection with fixed priorities (`X` $\rightarrow$ `Y` $\rightarrow$ `Z`) throughout training). Second, a scenario where priority orders switch at episode 1700 ($\Delta\pi_1$: `X` $\rightarrow$ `Y` $\rightarrow$ `Z` to `Y` $\rightarrow$ `Z` $\rightarrow$ `X`). Third, multiple priority changes with first at episode 1700 ($\Delta\pi_1$: `X` $\rightarrow$ `Y` $\rightarrow$ `Z` to `Y` $\rightarrow$ `Z` $\rightarrow$ `X`) and then at episode 3500 ($\Delta\pi_2$: `Y` $\rightarrow$ `Z` $\rightarrow$ `X` to `Z` $\rightarrow$ `X` $\rightarrow$ `Y`).

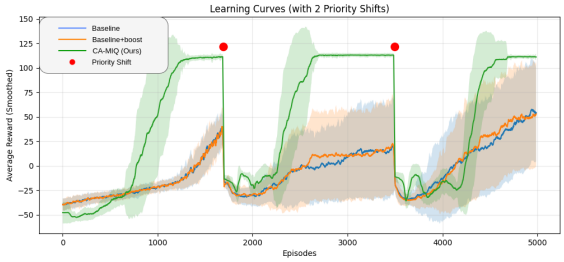
Hyper-parameters are shared across all agents: learning-rate $\alpha = 0.1$, discount factor $\gamma = 0.99$, initial exploration rate $\varepsilon_0 = 1.0$, minimum exploration rate $\varepsilon_{\min} = 0.1$, with linear decay over the full training horizon. The intrinsic reward is a weighted combination of novelty-based exploration (30%), information location awareness (40%), and



(a) *Static Priorities*



(b) *Single Priority Shift*



(c) *Multiple Priority Shifts*

Fig. 2: Comparison of reward trends under static and dynamic conditions.

priority alignment (30%). Boost factor $\lambda_{\text{boost}} = 2$, reset factor $\lambda = 0.5$. For each scenario, we ran 100 independent training sessions, each consisting of 5000 episodes. To ensure robustness, the agent is tested in randomly selected environments from a pool of five predefined configurations, each with different layouts for the target position, agent starting location, ditches, and information locations. All environments maintain the 4×4 grid structure and require collecting three pieces of information in the correct sequence. Performance metrics were averaged across these runs to ensure statistical significance. Key metrics included mission success rate, information collection success rate (i.e., the agent solved the task of collecting the necessary types of information), recovery success and time after priority changes (latter is measured in episodes).

V. Results & Discussion

A. Overall Performance Comparison

The experimental results demonstrate that CA-MIQ outperforms both baseline with and without boosting across all metrics. As shown in Table I, CA-MIQ achieved a mission success rate of 65.9% in the single priority shift scenario, more than three times higher than both standard baseline (18.5%) and boosted baseline (18.2%). Similarly, information collection success rates were substantially higher with CA-MIQ (69.2%) compared to the baselines (20.8% and 20.4%, respectively). The performance gap is also notable in the more challenging multiple priority shifts scenario (Table II), where CA-MIQ maintained a 50.2% mission success rate while both baseline approaches struggled with success rates below 15%. This demonstrates CA-MIQ's robustness when faced with multiple contextual changes, which better reflects the dynamic nature of real-world SAR operations.

B. Priority Adaptation Mechanisms

CA-MIQ's performance stems from two key mechanisms. First, transient epsilon boosting temporarily increases exploration after detecting a priority shift, helping the agent escape local optima established under previous priority orderings. Second, selective critic reset preserves general environment knowledge while targeting only value estimates related to information collection for relearning. This balanced approach allows CA-MIQ to discover new optimal policies efficiently while maintaining valuable knowledge about the environment's dynamics, explaining its sustained performance even after multiple priority shifts. Results show that CA-MIQ achieved a 100% recovery success rate in both single shift and multiple shifts scenarios, whereas both baseline approaches completely failed to adapt (0% recovery success). The average recovery time for CA-MIQ was 685 episodes after a single shift and 766 episodes for multiple shifts, showing consistent adaptation capabilities even in complex scenarios.

C. Learning Dynamics and Stability

The learning curves in Figure 2 provide further insights into the behavior of each approach. In the static environment (Figure 2a), all three approaches eventually converge to similar performance levels, though CA-MIQ learns more efficiently, reaching higher rewards earlier in training. In the single priority shift scenario (see Figure 2b), immediately after the priority change (marked by a red dot), all approaches experience a performance drop as expected. While the baseline methods struggle to recover, with their performance remaining below pre-shift levels, CA-MIQ rapidly adapts and returns to near-optimal performance levels. The multiple priority shifts scenario (Figure 2c) further emphasises this pattern. After each shifts, CA-MIQ experiences a temporary performance drop but consistently recovers successfully. The baseline approaches show minimal recovery after shifts, demonstrating CA-MIQ's stability and resilience to repeated contextual changes.

D. Ablation

The ablation study (Table III) provides empirical validation of each component's contribution to overall CA-MIQ's performance. The most critical finding is that removing state novelty has the most severe impact, reducing mission success by 31.5% and information collection by 17.2%. This suggests

that novelty-driven exploration is crucial for achieving long-term mission goals in dynamic environments. Removing priority alignment and awareness components resulted in a 5.8% decrease in mission success and 4.9% decrease in information collection, confirming that guiding exploration based on current priorities is essential for effective adaptation. Interestingly, while the exploration boost mechanism improved mission success by 2.0% and information collection by 1.4%, its impact was less dramatic than the novelty and priority alignment components. This suggests that the core intrinsic motivation components are more fundamental to CA-MIQ's success than the boost mechanism.

E. Limitations

The effectiveness of our selective reset mechanism depends on the structure of the SAR task and the nature of the encountered priority shifts. Currently, adaptation following a priority shift requires approximately 600–800 episodes, a number which might be reduced through adaptive reset scheduling or parameterised exploration boosts. Additionally, the fixed coefficients balancing novelty, information-location awareness, and priority alignment are effective in our SAR scenarios but lack generalisability. An adaptive weighting mechanism could automatically adjust these parameters based on mission characteristics and agent performance. System robustness depends on accurate and timely detection of context changes. Incorporating Bayesian-based detection or distribution-shift monitoring would improve reliability without requiring explicit signals. Furthermore, while our implementation avoids the computational overhead inherent to the ensemble-based uncertainty estimation employed in [3], it is inherently limited to discrete state–action spaces. Although we deliberately chose a simplified environment to match the discrete SAR setting and to isolate the contribution of priority-aligned intrinsic exploration, extending CA-MIQ to deep critics is straightforward (two heads with the same gate). We leave this engineering step and large-scale benchmarks to future work.

VI. Conclusion

We introduce CA-MIQ (Context-Aware Max-Information Q-learning) to address a critical challenge in RL-based systems: adapting to dynamic changes in the priority order of information types the agent is expected to focus on. The frameowrk enhances information-gathering strategies when mission objectives shift unexpectedly by integrating dual critics to encourage the reduction of priority-driven uncertainty with targeted exploration boosting and selective reset mechanisms. Our approach demonstrates up to 4× higher mission success rates compared to baselines and near-perfect recovery following priority shifts. The framework's strong performance in SAR-inspired simulations highlights its potential for broader application in domains requiring adaptive information gathering under evolving constraints. CA-MIQ effectively bridges the gap between algorithmic and priority-driven exploration, marking an important advancement for autonomous systems in dynamic, mission-critical environments.

References

[1] P. Ladosz, L. Weng, M. Kim, and H. Oh, "Exploration in deep reinforcement learning: A survey," *Information Fusion*, vol. 85, pp. 1–22, 2022.

[2] J.-H. Ewers, D. Anderson, and D. Thomson, "Deep reinforcement learning for time-critical wilderness search and rescue using drones," *Frontiers in Robotics and AI*, vol. 11, p. 1527095, 2025.

[3] B. Sukhija, S. Coros, A. Krause, P. Abbeel, and C. Sferrazza, "Maxinforl: Boosting exploration in reinforcement learning through information gain maximization," *arXiv preprint arXiv:2412.12098*, 2024.

[4] B. Hao and T. Lattimore, "Regret bounds for information-directed reinforcement learning," *Advances in neural information processing systems*, vol. 35, pp. 28 575–28 587, 2022.

[5] D. Panagopoulos, A. Perrusquia, and W. Guo, "Selective exploration and information gathering in search and rescue using hierarchical learning guided by natural language input," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2024, pp. 1175–1180.

[6] A. Rahman, A. Bhattacharya, T. Ramachandran, S. Mukherjee, H. Sharma, T. Fujimoto, and S. Chatterjee, "Adversar: Adversarial search and rescue via multi-agent reinforcement learning," in *2022 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, 2022, pp. 1–7.

[7] Y. Chen and J. Xiao, "Target search and navigation in heterogeneous robot systems with deep reinforcement learning," *Machine Intelligence Research*, vol. 22, no. 1, pp. 79–90, 2025.

[8] M. Zhou, X. Wang, C. Wang, and J. Wang, "Multi-robot cooperative target search based on distributed reinforcement learning method in 3d dynamic environments," *Drones and Autonomous Vehicles*, vol. 1, no. 4, p. 10012, 2024.

[9] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[10] N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu, "Boltzmann exploration done right," *Advances in neural information processing systems*, vol. 30, 2017.

[11] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," *Advances in neural information processing systems*, vol. 24, 2011.

[12] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *International conference on machine learning*. PMLR, 2017, pp. 2778–2787.

[13] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," *Advances in neural information processing systems*, vol. 29, 2016.

[14] J. Achiam and S. Sastry, "Surprise-based intrinsic motivation for deep reinforcement learning," *arXiv preprint arXiv:1703.01732*, 2017.

[15] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," *arXiv preprint arXiv:1810.12894*, 2018.

[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[17] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[18] A. Hallak, D. Di Castro, and S. Mannor, "Contextual markov decision processes," *arXiv preprint arXiv:1502.02259*, 2015.

[19] H. Ren, A. Sootla, T. Jafferjee, J. Shen, J. Wang, and H. Bou-Ammar, "Reinforcement learning in presence of discrete markovian context evolution," *arXiv preprint arXiv:2202.06557*, 2022.

[20] K. Lee, Y. Seo, S. Lee, H. Lee, and J. Shin, "Context-aware dynamics model for generalization in model-based reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5757–5766.

[21] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," *Operations Research*, vol. 66, no. 1, pp. 230–252, 2018.

[22] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.



https://dspace.org/

Learning what matters now: a dual-critic context-aware RL framework for priority-driven information gain

Panagopoulos, Dimitris; Perrusquía, Adolfo; Guo, Weisi

2025-10-05
Institute of Electrical and Electronics Engineers (IEEE)
https://doi.org/10.1109/smc58881.2025.11342671
Panagopoulos D, Perrusquía A, Guo W. (2025) Learning what matters now: a dual-critic context-aware RL framework for priority-driven information gain. In: Proceedings of the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 5-8 Oct 2025, Vienna, Austria, pp. 5655-5660

Downloaded from a DSpace repository. https://dspace.org/