Selective Amnesia: Observation Unlearning in Reinforcement Learning

Extended Abstract

Yue Yang†*
Monash University
Melbourne, Australia
yue.yang1@monash.edu

Jinhao Li*
Monash University
Melbourne, Australia
jinhao.li@monash.edu

Hao Wang†
Monash University
Melbourne, Australia
hao.wang2@monash.edu

ABSTRACT

Although the concept of machine unlearning has been widely explored in the past few years, unlearning in reinforcement learning (RL) models remains underdeveloped. In this paper, we undertake an in-depth exploration of reinforcement unlearning (RUL), a novel and challenging concept within the field of RL and machine unlearning. We investigate the inherent difficulties associated with RUL, pinpointing two critical factors that contribute to its complexity: agent-environment interactions and the sequential nature of decision-making. To tackle these challenges, we propose an unlearning algorithm that addresses the fundamentals of RUL from the perspective of environment observations, enabling observation-level unlearning for both tabular and deep Q-learning. By quantitatively assessing the effects of observations through state-action values and modifying and retracing the policy trajectories established by the original model, we demonstrate that, under reasonable assumptions, RUL can effectively eliminate both the immediate and subsequent impacts of the targeted unlearning observation. Empirical evaluations also validate the effectiveness of our RUL approach.

KEYWORDS

Reinforcement learning, machine unlearning

ACM Reference Format:
Yue Yang†*, Jinhao Li*, and Hao Wang†. 2026. Selective Amnesia: Observation Unlearning in Reinforcement Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/SOOJ3168

1 INTRODUCTION

Reinforcement learning (RL) has driven major AI advances, from fine-tuning large language models [8] to recommender systems [1]. As RL deployment grows, *reinforcement unlearning*—removing unsafe or undesired knowledge from trained agents (e.g., faulty sensor-induced driving behaviours or leaked confidential visuals) [2, 4, 5]—becomes increasingly important, yet systematic methods remain limited [7, 9]. Although one might frame unlearning by modifying MDP components (state, reward, transitions), doing so effectively alters the environment (e.g., suppressing rewards for specific state–action pairs), contradicting the goal of unlearning the *policy* while keeping the environment intact. This paper explores an unlearning approach at the observation level, where specific observations serve as the unlearning targets. We introduce ***RUL*** to denote reinforcement unlearning. We begin by focusing on unlearning models trained through *Q-learning*. To be specific, we first pinpoint the Q-value affected by the observation targeted for unlearning and then eradicate its influence across the entire action policy landscape. By quantifying the contributions of the targeted observation through Q-values using an overwriting method (termed ***Q-Cover***) and backtracking all policy trajectories of the original model (termed ***Twin Q-Backward***), we demonstrate that, under mild assumptions, our proposed RUL approach can comprehensively eliminate both the direct and lingering impacts of the observation while preserving the contributions from the remaining environment in tabular Q-learning. Our thorough experimental evaluations confirm the effectiveness of our RUL.

2 REINFORCEMENT UNLEARNING

Let o_i denote the targeted observation and $\mathbf{o}_{-i}$ be the aggregation of remaining observations. We aim to unlearn o_i such that for any given initial value of an unlearned observation, in the original MDP, expressed as $\forall\, o_{i,t} \in \mathbf{o}_t\big|_{t=0} \in O :\ o_{i,t} = o_i^{\text{def}}$, where o_i^{def} is the initial value of o_i provided by the environment. Here, we provide the definition of our RUL in the context of unlearning o_i from the RL model trained by Q-learning. The model can be expressed in terms of observation as $\pi^*(\mathbf{o}) = \text{argmax}_{\mathbf{a}\in\mathcal{A}}\, Q^*(\mathbf{o}, \mathbf{a})$, where $\mathbf{o} = o_i \cup \mathbf{o}_{-i} \in O$. The action policy after applying RUL is denoted as $\pi^{\text{UL}}(\mathbf{o}_{-i} \cup o_i)$. As the influence of the targeted observation o_i has been fully removed, its value no longer affects the outcomes of the revised action policy, which can then be rewritten as $\pi^{\text{UL}}(\mathbf{o}_{-i})$, where $\mathbf{o}_{-i} \in O^{\text{PO}}$. The Q-value after RUL is denoted as Q^{UL}.

Definition 1 (*Reinforcement Unlearning*). *Our RUL has two essential properties: 1) the Q-value after RUL complies with the following Bellman equation:* $\forall \mathbf{o}_{-i} \in O^{\text{PO}}$*:*

$$Q^{\text{UL}}(\mathbf{o}_{-i}, \pi^{\text{UL}}(\mathbf{o}_{-i})) = \mathbb{E}[r + \gamma Q^{\text{UL}}(\mathbf{o}'_{-i}, \pi^{\text{UL}}(\mathbf{o}'_{-i}))] \tag{1}$$

and 2) the action outcomes of the revised action policy should be identical to the action policy trained from scratch without the removed observation. The latter action policy is denoted as π^{RT}*. The second property can be formulated as*

$$\pi^{\text{UL}}(\mathbf{o}_{-i}) = \pi^{\text{RT}}(\mathbf{o}_{-i}), \quad \forall \mathbf{o}_{-i} \in O^{\text{PO}}. \tag{2}$$

* Equal Contribution.
†Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/SOOJ3168

We aim to unlearn the influence of the targeted observation o_i. Since o_i starts with a initial given value of o_i^{def}. When the value of o_i changes on a specific trajectory generated by the Q-table, we can then identify that the targeted observation may influence the agent's decision-making. Assume the value of o_i changes at time step t, i.e., $o_{i,t} \neq o_i^{\text{def}}$.

Q-Cover: To remove such influence, we first perform the CV step, using the Q-value of $Q^*(\mathbf{o}_{-i,t} \cup o_i^{\text{def}}, \pi^*(\mathbf{o}_{-i,t} \cup o_i^{\text{def}}))$ to replace the current Q-value of $Q^*(\mathbf{o}_{-i,t} \cup o_{i,t}, \pi^*(\mathbf{o}_{-i,t} \cup o_{i,t}))$. The revised Q-value is then denoted as Q^{CV}. The CV step can be summarized as: $\forall \mathbf{o}_{-i,t} \in \mathbf{o}_t, (\mathbf{o}_t, \mathbf{a}_t) \in \tau, \mathbf{a}_t = \pi^*(\mathbf{o}_{-i,t} \cup o_{i,t}), \tau \in \mathcal{L}$:

$$Q^{\text{CV}}(\mathbf{o}_{-i,t} \cup o_{i,t}, \mathbf{a}_t) := Q^*(\mathbf{o}_{-i,t} \cup o_{i,t}, \mathbf{a}_t) = Q^*(\mathbf{o}_{-i,t} \cup o_i^{\text{def}}, \mathbf{a}_t). \tag{3}$$

Note that only conducting the CV step does not suffice the requirement of RUL, as the Q-value of the current observation-action pair also influences all preceding trajectories that have led to the current state. Consequently, we develop a backward update process along the recorded trajectory, referred to as the BS step.

Q-Backward Substitution: Starting from the same observation-action pair in the CV step, we iteratively update the Q-values using the Bellman equation and move backward toward the start of the trajectory. The updated Q-value is denoted as Q^{BS}. For the most recent observation-action pair $(\mathbf{o}_{t-1}, \mathbf{a}_{t-1})$, the BS step can be formulated as $\forall (\mathbf{o}_{t-1}, \mathbf{a}_{t-1}), (\mathbf{o}_t, \mathbf{a}_t) \in \tau, \mathbf{a}_{t-1} = \pi^*(\mathbf{o}_{-i,t-1} \cup o_i^{\text{def}}), \tau \in \mathcal{L}$:

$$Q^{\text{BS}}(\mathbf{o}_{t-1}, \mathbf{a}_{t-1}) := Q^*(\mathbf{o}_{t-1}, \mathbf{a}_{t-1}) = r_{t-1} + \gamma Q^{\text{CV}}(\mathbf{o}_{-i,t} \cup o_{i,t}, \mathbf{a}_t). \tag{4}$$

We repeat the above operation till the observation-action pair at the first (or initial) time step of the trajectory τ. After processing all trajectories in the trajectory set $\mathcal{L}$, we can conclude that for any time step t and $\forall \mathbf{o}_{-i,t} \in \mathbf{o}_t, \mathbf{o}_{-i,t+1} \in \mathbf{o}_{t+1}, \mathbf{o}_t, \mathbf{o}_{t+1} \in O$, we have $Q^{\text{BS}}(\mathbf{o}_{-i,t}, \mathbf{a}_t) = r_t + \gamma Q^{\text{BS}}(\mathbf{o}_{-i,t+1}, \mathbf{a}_{t+1})$.

Q-Backward Optimization: After the elimination of all the effects of the observation we intended to remove, we now apply the Bellman optimal equation to the same trajectory set $\mathcal{L}$, aiming to derive the final action policy. Subsequently, the action policy after RUL can be formulated as $\pi^{\text{UL}}(\mathbf{o}_{-i}) = \text{argmax}_{\mathbf{a} \in \mathcal{A}} Q^{\text{UL}}(\mathbf{o}_{-i}, \mathbf{a}_t)$. Finally, the revised Q-table after RUL aligns with the first property of our RUL definition defined in Eq. (1). The corresponding algorithmic procedure is detailed in Algorithm 1. In the experiments presented in Section 3, we validate the second property of RUL using the revised Q-table, which is compared to the Q-table trained from scratch without the removed observation.

3 EXPERIMENTS AND EVALUATION

We evaluate observation-space RL unlearning by defining Unlearning Accuracy (UAcc) as the action agreement between the unlearned policy and a scratch-trained policy that never saw the removed observation, measured under the same observations: $\text{UAcc} = \frac{1}{|O_{\text{eval}}|} \sum_{k=1}^{|O_{\text{eval}}|} \mathbb{I}\left[\pi^{\text{UL}}(\mathbf{o}_{-i,k}) = \pi^{\text{RT}}(\mathbf{o}_{-i,k})\right]$, where O_{eval} is a set of observations for unlearning evaluation and $\mathbb{I}(\cdot)$ is an indicator which is equal to 1 only when the two models output the same action. Otherwise, this indicator is equal to zero.

The tabular Q-unlearning experiments are conducted on two standard testbeds: the LINKT Chain environment [6] and *Cliff Walking* from OpenAI Gym [3]. In LINKT Chain, the agent starts at a fixed position on a 1D chain and chooses actions {*left*, *right*}. To complete the task, it must first move to the left end to obtain a key, then travel to the right end to unlock a door. The observation is a 2-tuple (position, has_key), where position $\in \{1, \ldots, 6\}$ indexes the chain from left to right. If the position information is deemed sensitive and must be unlearned, applying our RUL to the Q-table yields an unlearned policy that is identical to a policy trained from scratch without the sensitive observation (move left before obtaining the key, and right afterwards), achieving unlearning accuracy = 1.0. In *Cliff Walking*, the agent navigates a grid from a start to a goal while avoiding the "cliff" along one edge, which incurs large penalties; the objective is to reach the goal efficiently while minimising the risk of falling. The observation is the agent's location, and actions are {*up*, *right*, *down*, *left*}. To create an observation-level unlearning target, we augment the observation with an additional indicator that records whether the agent has visited a designated waypoint, incentivising the agent to pass through that point before reaching the goal. After unlearning this auxiliary observation with our RUL, the mean unlearning accuracy across ten independently trained policies is 85.76% with standard deviation 4.56%. Beyond these two environments, we study a 3×3 gridworld to further evaluate RUL for tabular Q-unlearning. In the original setting, the agent observes its current position and a Boolean flag indicating whether it has collected a token. When the token has not yet been collected, the agent may either detour to collect it or proceed directly to the exit. We consider two unlearning tasks. In Task 1, we remove the agent's ability to observe the token status (while the token reward of 2 remains), so the agent no longer knows whether the token has been collected. In Task 2, we additionally remove the token-collection reward, making the optimal policy to head directly to the exit via the shortest path from any location. The unlearned and scratch-trained Q-tables match, again yielding 100% unlearning accuracy.

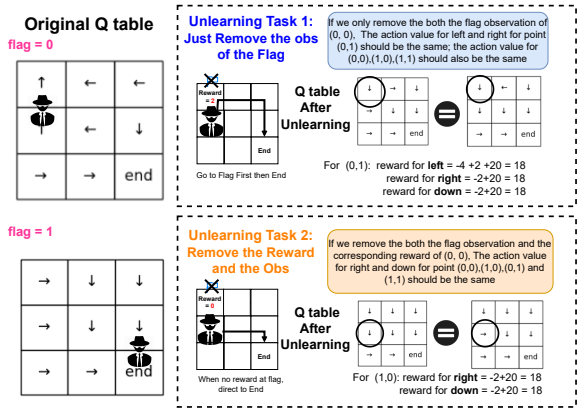

Figure 1: g Q-table unlearning under two task settings

4 CONCLUSION

In this paper, we study *state-level* reinforcement unlearning, formalise the problem setting, and propose **RUL** for Q-learning, which provably and empirically removes targeted information by retracing and editing the original policy trajectory, while largely preserving non-targeted behaviours and previously learned knowledge.

REFERENCES

[1] M. Mehdi Afsar, Trafford Crump, and Behrouz Far. 2022. Reinforcement Learning based Recommender Systems: A Survey. *ACM Comput. Surv.* 55, 7, Article 145 (dec 2022), 38 pages. https://doi.org/10.1145/3543846

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine Unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 141–159. https://doi.org/10.1109/SP40001.2021.00019

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. arXiv:arXiv:1606.01540

[4] Jonathan Brophy and Daniel Lowd. 2021. Machine unlearning for random forests. In *International Conference on Machine Learning*. PMLR, 1092–1104.

[5] Yinzhi Cao, Alexander Fangxiao Yu, Andrew Aday, Eric Stahl, Jon Merwine, and Junfeng Yang. 2018. Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (Incheon, Republic of Korea) *(ASIACCS '18)*. Association for Computing Machinery, New York, NY, USA, 735–747. https://doi.org/10.1145/3196494.3196517

[6] Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. 2022. Exploration-Guided Reward Shaping for Reinforcement Learning under Sparse Rewards. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 5829–5842. https://proceedings.neurips.cc/paper_files/paper/2022/file/266c0f191b04cbbbe529016d0edc847e-Paper-Conference.pdf

[7] Romain Laroche and Remi Tachet Des Combes. 2022. Beyond the Policy Gradient Theorem for Efficient Policy Updates in Actor-Critic Algorithms. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, 5658–5688. https://proceedings.mlr.press/v151/laroche22a.html

[8] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024. Data-efficient Fine-tuning for LLM-based Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 365–374. https://doi.org/10.1145/3626772.3657807

[9] Dayong Ye, Tianqing Zhu, Congcong Zhu, Derui Wang, Zewei Shi, Sheng Shen, Wanlei Zhou, and Minhui Xue. 2024. Reinforcement Unlearning. https://arxiv.org/abs/2312.15910. arXiv:2312.15910 [cs.CR]