Learning to Undo: Rollback-Augmented Reinforcement Learning with Reversibility Signals

Andrejs Sirstkins[1☯¤*], Omer Tariq[2☯*], Muhammad Bilal[1‡*],

1 School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, United Kingdom
2 Newbility, 2F 115 (04768) Wangsimni-ro, Seongdong-gu, Seoul, South Korea

¤Lead author
☯These authors contributed equally to this work.
‡Supervisor, editor and steering professor
*info@benarktech.co.uk; omertariq@kaist.ac.kr; m.bilal8@lancaster.ac.uk;

Abstract

This paper postulates a novel reversible learning framework designed to enhance the robustness and efficiency of value-based Reinforcement Learning (RL) agents, specifically addressing their pervasive vulnerability to value overestimation and instability in partially irreversible environments. The framework instantiates two complementary core mechanisms: an empirically derived transition reversibility measure ($\Phi(s, a)$) and a selective state-rollback operation. To achieve this, we introduce an online, per-state-action estimator (Φ) that quantifies the likelihood of returning to a prior state within a fixed horizon K. This measure is used to adjust the penalty term during temporal difference updates dynamically, integrating reversibility awareness directly into the value function. Crucially, the system incorporates a selective rollback operator: when an action yields an expected return markedly lower than its instantaneous estimated value (violating a predefined threshold), the agent is penalized and reverts to the preceding state rather than progressing. This strategically interrupts sub-optimal, high-risk trajectories and avoids catastrophic steps. By synergistically combining this reversibility-aware evaluation with targeted rollback, the proposed methodology demonstrably improves safety, performance, and stability. Empirically, in the CliffWalking-v0 domain, the framework reduced catastrophic falls by over 99.8% and yielded a 55% increase in mean episode return. Similarly, in the Taxi-v3 domain, it suppressed illegal actions by $\geq 99.9\%$ and achieved a 65.7% improvement in cumulative reward, while also sharply reducing reward variance in both environments. Ablation studies confirm the rollback mechanism is the critical component underlying these substantial safety and performance gains, marking a robust step toward safe and reliable sequential decision-making.

1 Introduction

Reinforcement learning (RL) paradigms have demonstrated state-of-the-art efficacy across an array of domains, from strategic board games such as Go and discrete control benchmarks like Atari, to real-world control problems in high-dimensional robotics and complex, unstructured environments [1]. The remarkable performance gains achieved by deep RL methods-through innovations in function approximation, experience replay, and actor–critic architectures-have reignited interest in deploying such algorithms for real-world decision-making tasks. Nevertheless, the transition of RL algorithms from controlled experimental settings to operational environments is frequently impeded by training-induced instability, sample inefficiency, and emergent unsafe behaviors [2]. A

primary factor contributing to these challenges is the pervasive overestimation of action-value functions [3], which skews policy improvement towards trajectories with spuriously optimistic reward predictions. When an agent's Q-function becomes biased towards overly optimistic reward estimates [4], it preferentially pursues statistically spurious or low-probability trajectories, precipitating oscillatory policy updates, prolonged convergence times, and, in worst-case scenarios, catastrophic failures within safety-critical infrastructures such as aerospace control systems or nuclear facility management.

Reversibility, an intrinsic aspect of human cognitive architectures, underpins our capacity for deliberative decision-making and adaptive learning. Individuals habitually assess not only the immediate reward associated with a given action, but also the extent to which that action can be reversed or counteracted by subsequent steps. This involves generating internal counterfactual simulations-mental "rollbacks"-to evaluate potential failure modes and to hedge against irreversible outcomes. Such meta-cognitive processes enable humans to engage in risk-sensitive exploration, to remediate mistakes via corrective maneuvers, and to maintain a consistent trajectory towards long-term objectives. Despite its foundational relevance to algorithmic safety and robustness, this latent human impulse to "undo" suboptimal decisions-and thereby explore alternative strategies without irrevocable consequence-remains scarcely addressed in existing RL research frameworks.

Embedding reversibility into an RL framework offers an illustrative principle for a broad spectrum of safety-critical applications [5–10]. Consider autonomous vehicular control, where irreversible errors-such as collisions-can precipitate loss of life or property damage; or robotic surgical assistants, where miscalibrated manipulations must be promptly retracted to avoid patient harm. Similarly, adaptive medical treatment planning algorithms must be able to backtrack from harmful dosage adjustments, and industrial process control systems must swiftly revert hazardous state transitions to prevent environmental or infrastructural compromise. In these contexts, the inability to retract or attenuate deleterious transitions can incur unacceptable risk. We address this exigency by integrating an online reversibility estimator-a learned function that predicts the probability of returning to a safe state distribution from any given transition-with an explicit rollback operator. Upon detection of high-risk transitions-quantified via this reversibility metric-the system effectuates a corrective "U-turn," restoring the agent to a prior checkpointed state. This mechanism not only constrains exploratory risk and prevents agent entrapment in irreversible error states but also attenuates policy divergence, thus facilitating stable convergence under rigorous safety constraints.

Conventional cures for Q-overestimation-dual/twin critics, bias-corrected evaluation, and conservative Q-learning-often trade accuracy for added critics, tighter update rules, and cautious behavior, inflating compute and sample cost [11]. In reversibility-aware RL, [12] learn a "precedence" score from raw trajectories to avoid irreversible regions, but their approach trains a Siamese classifier tied to the behavior policy, uses a fixed temporal window, relies on a global threshold to gate actions, and never actually undoes a damaging step. We address these gaps with a rollback-augmented framework that couples (i) a per-state–action empirical reversibility estimator $\Phi(s, a)$, computed online via a FIFO return-within-K test and updated by a light EMA, with (ii) an explicit "U-turn" rollback that fires only when the reversibility-penalized TD target falls below a threshold; Φ also induces a localized penalty in the TD update. This design eliminates the need for a learned Siamese model, adapts naturally to different horizons via K, replaces a blunt global irreversibility proxy with per-state–action estimates, and, crucially, equips the agent with an actionable undo. Empirically, in `CliffWalking-v0` we cut catastrophic falls by $\geq 99.8\%$ and improve mean return by $\sim 55\%$ while collapsing return variance by $\sim 71\%$; in `Taxi-v3` we suppress illegal actions by $\geq 99.9\%$

with $\sim 66\%$ return gains and $\sim 59\%$ variance reduction. Our contributions are: (1) a scalable, model-free, per-state–action reversibility estimator that avoids classifier training; (2) an explicit rollback operator integrated into tabular Q-learning and SARSA updates; (3) a principled coupling of Φ-shaping and selective rollback that bounds downside without choking exploration; and (4) extensive evaluation, sensitivity analyses, and ablations that isolate which components matter for safety and performance.

2 Background

Reinforcement learning has made incremental improvements over the last few decades. Overestimation of action values has long been recognized as a key obstacle to stable and efficient learning in value-based RL, making it one of the main challenges to address. In this section, we provide a brief overview of the key foundational concepts.

2.1 Reinforcement Learning and Markov Decision Processes

Reinforcement learning (RL) frames sequential decision-making as a Markov decision process (MDP) [1]

$$(\mathcal{S}, \mathcal{A}, P, R, \gamma), \tag{1}$$

where an agent in state $s \in \mathcal{S}$ chooses action $a \in \mathcal{A}$, receives reward r, and transitions to $s' \sim P(\cdot \mid s, a)$ with discounted return

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}. \tag{2}$$

Value-based RL approximates the action-value function

$$Q(s, a) \approx \mathbb{E}[\, G_t \mid s_t = s, a_t = a \,] \tag{3}$$

via temporal-difference updates.

2.2 Tabular Q-Learning

Q-learning [13] updates a table of values via

$$Q(s, a) \leftarrow Q(s, a) + \alpha\big(r + \gamma \max_{a'} Q(s', a') - Q(s, a)\big). \tag{4}$$

This off-policy rule can converge to the optimal action-value function under sufficient exploration.

2.3 SARSA

SARSA is on-policy: it updates toward the value of the action actually taken, sampling $a' \sim \pi(\cdot \mid s')$:

$$Q(s, a) \leftarrow Q(s, a) + \alpha\big(r + \gamma Q(s', a') - Q(s, a)\big). \tag{5}$$

This ensures updates remain consistent with the agent's current policy.

2.4 Selection and Evaluation

Early work tackled this by decoupling selection and evaluation: Double Q-Learning [14] maintains two independent estimators-using one to choose actions and the other to evaluate them-curbing maximization bias in both tabular and deep settings [4]. TD3 extends this idea to continuous control by clipping between twin critics and delaying policy updates to further suppress overoptimism [15]. Rather than relying solely on multiple networks, Maxmin Q-Learning maintains N critics and interpolates between their highest and lowest predictions via a tunable parameter κ. By adjusting κ, it trades off optimism against conservatism, yielding tighter theoretical bounds on estimation error and empirical gains across benchmarks [11].

2.5 Precedence Estimation

As introduced by Grinsztajn N. [12], *precedence* is a self-supervised statistic capturing the temporal "direction" between two states under a fixed policy π and horizon T. They define

$$\psi_{\pi,T}(s,s') = \frac{\mathbb{E}_{\tau\sim\pi}\big[\,|\{(t,t') : t' < t,\ s_t = s,\ s_{t'} = s'\}|\big]}{\mathbb{E}_{\tau\sim\pi}\big[\,|\{(t,t') : t \neq t',\ s_t = s,\ s_{t'} = s'\}|\big]}, \tag{6}$$

estimating the probability that state s appears after s' in trajectories of length $\leq T$. In practice, one samples trajectories, collects state-pairs within a window $|t - t'| \leq w$, and computes the fraction with $t' < t$:

- If $\psi \approx 1$, transitions $s \to s'$ are essentially irreversible.
- If $\psi \approx 0.5$, no consistent ordering exists, indicating reversibility.

They then lift ψ to an action-level score by averaging over next-state distributions:

$$\bar{\phi}_\pi(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\big[\psi_{\pi,T}(s',s)\big], \tag{7}$$

which serves as a data-driven proxy for reversibility without external labels or models.

2.6 Related Work and Comparative Positioning

Although these approaches each mitigate overestimation in different ways-through alternate estimators, bias–variance blending, or learned state–action reversibility they stop short of explicitly undoing poor decisions. Safe exploration approaches in MDPs [16] similarly aim to avoid irreversible failures, but do not provide rollback mechanisms.

Safe exploration in RL has been widely studied due to the risks of unsafe behavior during training and deployment. Existing approaches can be grouped into three broad categories: (i) *constraint-based formulations*, (ii) *verification-based methods*, and (iii) *optimization-based trade-off techniques*.

2.6.1 Constraint-based safe exploration

Wachi et al. [17] introduce the *Generalized Safe Exploration (GSE)* framework, which unifies common safe RL formulations-cumulative, state, and instantaneous constraints-into a meta-algorithm (MASE) with high-probability safety guarantees. By penalizing unsafe actions before actual violations, MASE ensures safety even during training, extending beyond average-case constraint satisfaction. Similarly, As et al. [18] propose *ActSafe*, a model-based approach that learns probabilistic dynamics models and couples optimistic exploration with pessimistic safety constraints. ActSafe provides finite-sample complexity guarantees while scaling to high-dimensional deep RL settings.

2.6.2 Verification-based safe RL.

Formal verification methods have also been applied to ensure provable safety during exploration. Wang and Zhu [19] propose *VELM* (Verified Exploration through Learned Models), which learns symbolic environment models amenable to reachability analysis. VELM constructs a shielding mechanism that confines the agent's actions to formally verified safe regions, thereby reducing violations without degrading reward performance. While powerful, such approaches often depend on the tractability of symbolic regression or approximations of nonlinear dynamics, which can limit applicability in highly stochastic or large-scale domains.

2.6.3 Reward–safety trade-off optimization.

Another line of research emphasizes balancing safety constraints with performance. Gu et al. [20] highlight the intrinsic gradient conflict between reward maximization and safety optimization. Their framework introduces gradient manipulation techniques to reconcile these conflicts, producing improved trade-offs across Safety-MuJoCo and OmniSafe benchmarks. This direction complements earlier constrained optimization methods (e.g., CPO [21], PPO-Lagrangian [22]), but with a sharper focus on handling conflicting optimization signals.

2.6.4 Positioning of this work.

In contrast to prior approaches that either enforce hard constraints (e.g., GSE, VELM) or resolve gradient conflicts [20], our work introduces a *reversibility-driven perspective*. We propose an empirical reversibility estimator coupled with a rollback operator that enables the agent not only to avoid unsafe regions but to actively *undo* detrimental steps.(Algorithm 1).This mechanism provides an additional layer of resilience absent in most existing safe exploration frameworks, which typically rely on forward-looking predictions or static safety filters. Unlike ActSafe, which guarantees safety by conservative set expansion, our rollback mechanism offers *dynamic recoverability*, making exploration less brittle in environments where occasional missteps are unavoidable. Moreover, by empirically demonstrating over 99% reduction in catastrophic actions and consistent return improvements, our method complements existing safe RL approaches by offering a pragmatic, model-free safeguard against irreversible outcomes.

Algorithm 1 Modified Q-Learning with Precedence and rollback

Require: $Q[s,a] \leftarrow Q_0,\ \Phi[s,a] \leftarrow \phi_0,\ \text{buffer} \leftarrow \emptyset,\ t \leftarrow 0$

while true **do**
  $a \leftarrow \epsilon\text{-}\texttt{greedy}(Q[s,\cdot])$
  observe reward r, next state s', and flag *done*
  $t \leftarrow t+1$
  for all records (s_0, a_0, d) in buffer **do**
    if $s' = s_0$ **then**
      $y \leftarrow 1$
    else if $t > d$ **then**
      $y \leftarrow 0$
    else
      continue
    end if
    $\Phi[s_0,a_0] \leftarrow (1-\alpha_\phi)\,\Phi[s_0,a_0] + \alpha_\phi\, y$
    remove (s_0, a_0, d) from buffer
  end for
  append $(s, a, t+K)$ to buffer
  $r' \leftarrow r - \lambda\left(1 - \Phi[s,a]\right)$
  if *done* **then**
    $\mathit{target} \leftarrow r'$
  else
    $\mathit{target} \leftarrow r' + \gamma\, \max_{a'} Q[s',a']$
  end if
  $\delta \leftarrow \mathit{target} - Q[s,a]$
  if $\mathit{target} \le T \cdot Q[s,a]$ **then**
    $\beta' \leftarrow \beta,\ \text{rollback} \leftarrow \text{true}$
  else
    $\beta' \leftarrow 1,\ \text{rollback} \leftarrow \text{false}$
  end if
  $Q[s,a] \leftarrow Q[s,a] + \alpha\,\beta'\,\delta$
  if rollback **and** $\neg$ *done* **then**
    $s \leftarrow s$
  else
    $s \leftarrow s'$
  end if
end while

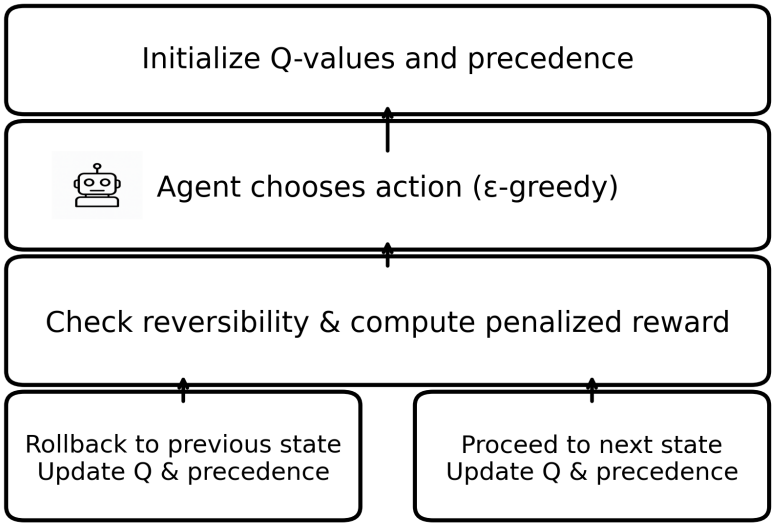

Fig 1. Reversible Q-learning with rollback and precedence.

3 System Overview

Precedence-based reversibility [12] offers a self-supervised signal for whether transitions can "undo" themselves, yet it exhibits four coupled weaknesses in practice. First, the Siamese classifier is trained purely on the agent's own trajectories and can overfit to policy-specific quirks; if the behavior policy is near-deterministic or fails to revisit certain state–action pairs, the estimator may systematically mislabel reversible transitions as irreversible (and vice versa). Second, a fixed temporal window w forces short-horizon judgments and obscures longer-range reversibility that requires extended return paths. Third, both the Reversibility-Aware Explorer (RAE) and Controller (RAC) rely on a single global threshold β to gate actions, a blunt control that struggles in heterogeneous state–action spaces and requires environment-specific retuning. Fourth, neither RAE nor RAC provides an explicit *rollback* mechanism; once a damaging move is taken, the agent cannot immediately undo it, leaving learning exposed when some irreversible steps are unavoidable. Related work on skill discovery has also implicitly leveraged reversibility, for example in unsupervised RL approaches such as DIAYN [23], where diversity-enforcing objectives yield reusable and often reversible behaviors. However, these works do not provide explicit rollback mechanisms.

Our approach. We replace classifier-based precedence with a lightweight, empirical reversibility estimate maintained online and coupled to an explicit rollback operator (Fig 1). Rather than fitting a Siamese model to policy-induced data, we enqueue each observed transition into a fixed-size FIFO structure of length K and update a per-state–action estimate $\Phi(s, a)$ via an exponential moving average. The horizon is thus controlled solely by K, allowing short- or long-range reversibility without retraining. We integrate Φ into temporal-difference (TD) learning through a localized penalty that is applied only when a *reversibility-penalized* target breaches a threshold. Because Φ is defined per (s, a), this yields fine-grained, data-driven shaping rather than a single global cutoff. Crucially, when the same threshold condition is violated, we execute an explicit rollback that resets the agent to the previous state, preventing

irreversible missteps from contaminating subsequent learning. This corrective principle is reminiscent of off-policy correction methods such as $Q(\lambda)$ with importance adjustments [24], but in contrast, our rollback mechanism directly intervenes at the state-transition level rather than adjusting the weighting of returns.

3.1 Empirical Reversibility via a Precedence FIFO

Consider a transition $(s_t, a_t) \to s_{t+1}$. Immediately after observation, we push a *pending record* (s_0, a_0, d) onto a FIFO list L, where $s_0 = s_t$, $a_0 = a_t$, and $d = t + K$ is a deadline by which a return to s_0 must occur to be counted as reversible. On each subsequent step, we scan pending records in L: (i) if the current state matches any s_0 before its deadline, we set $y = 1$ and remove that record; (ii) if the deadline is exceeded without a match, we set $y = 0$ and remove it; (iii) otherwise, the record remains pending. Because each record is dequeued no later than K steps after insertion, $|L| \leq K$ and memory is bounded.

When a record resolves with label $y \in \{0, 1\}$, we update the reversibility table $\Phi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ by an exponential moving average (EMA):

$$\Phi[s_0, a_0] \leftarrow (1 - \alpha_\phi)\, \Phi[s_0, a_0] + \alpha_\phi\, y, \tag{8}$$

with small learning rate $\alpha_\phi \ll 1$. Under stationarity and sufficient visitation, the EMA converges to the probability of returning to s_0 within K steps. Intuitively, frequent returns drive $\Phi \to 1$ (high reversibility), whereas persistent non-returns drive $\Phi \to 0$ (high irreversibility). Initialization of Φ encodes prior risk posture: pessimistic ($\Phi \approx 0$), neutral ($\Phi \approx 0.5$), or optimistic ($\Phi \approx 1$); we study these priors empirically to simulate different exploration biases.

3.2 TD Learning with Penalization and Rollback

We maintain two tabular objects: the action-value function $Q[s, a]$ and the reversibility estimate $\Phi[s, a]$. At each step, we form a *penalized reward*

$$r' \;=\; r \;-\; \lambda \big(1 - \Phi[s_t, a_t]\big), \tag{9}$$

where $\lambda \geq 0$ scales the irreversibility penalty. This yields the modified TD error for Q-learning

$$\delta \;=\; r' + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t), \tag{10}$$

and for SARSA

$$\delta \;=\; r' + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t). \tag{11}$$

We introduce a multiplicative factor β to amplify corrections when the (unpenalized) target underperforms the current estimate by more than a threshold $T \in (0, \infty]$:

$$\beta \;=\; \begin{cases} P, & \text{if } r + \gamma \max_{a'} Q(s_{t+1}, a') \leq T\, Q(s_t, a_t) \quad \text{(Q-learning)}, \\ P, & \text{if } r + \gamma Q(s_{t+1}, a_{t+1}) \leq T\, Q(s_t, a_t) \quad \text{(SARSA)}, \\ 1, & \text{otherwise}, \end{cases} \tag{12}$$

with $P \in (0, \infty]$ the penalty level used only in adverse targets. The value update is then

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\, \beta\, \delta. \tag{13}$$

Rollback operator. When the threshold condition in Eq (12) is triggered, we execute a rollback by setting the next state to the current state (and, for SARSA, the next action to the current action). For Q-learning, the rollback operator is defined as

$$s_{\text{next}} = \begin{cases} s_t, & \text{if the threshold is violated,} \\ s_{t+1}, & \text{otherwise.} \end{cases} \tag{14}$$

For SARSA, the operator extends to both state and action:

$$(s_{\text{next}}, a_{\text{next}}) = \begin{cases} (s_t, a_t), & \text{if the threshold is violated,} \\ (s_{t+1}, a_{t+1}), & \text{otherwise.} \end{cases} \tag{15}$$

This explicit "U-turn" prevents low-quality, potentially irreversible transitions from propagating errors and stabilizes exploration under risk.

3.3 Design Rationale and Behavioral Control

The FIFO construction bounds memory and enforces a clear K-step notion of reversibility; increasing K models higher "patience" before declaring a transition irreversible. The per-state–action Φ produces localized penalties via Eq (9), in contrast to a global β cutoff in prior work; this improves compatibility with heterogeneous state–action topologies. The thresholded scaling β in Eq (12) sharpens corrective updates only when warranted, avoiding chronic pessimism. Finally, the rollback in Eq (14) and Eq (15) adds an *actionable* recovery primitive absent from precedence-only schemes, reducing contamination from catastrophic steps. Together, these components yield a single, continuous process that neutralizes the four weaknesses of precedence-based reversibility without heavyweight classifiers or environment models.

3.4 Interpreting Hyperparameters

The horizon K governs the reversibility granularity and indirectly the agent's tolerance for delayed recovery; smaller K yields conservative, short-horizon judgments, while larger K captures longer detours. The initialization of Φ encodes prior risk appetite (pessimistic, neutral, optimistic) and can be selected to match domain priors or safety requirements. The threshold T controls the acceptance level before rollback: higher T triggers rollbacks sooner (safer but potentially slower learning), while lower T tolerates temporary degradation to preserve exploration. We study sensitivity to $(K, \lambda, T, P, \alpha_\phi)$ in Section 5.

4 Simulation

4.1 Environments

All experiments were conducted using Gymnasium v1.2.0 (Farama Foundation, 2025)[1]. This framework extends the original OpenAI Gym API [25], which remains a standard benchmark suite for reproducible reinforcement learning research. While our study focuses on single-agent tabular domains, similar reproducibility concerns have motivated the development of multi-agent environments such as PettingZoo [26], which extends the Gym interface to multi-agent RL. Two canonical tabular "toy-text" domains were chosen to evaluate the reversible-RL algorithm under diverse yet tractable conditions:

[1] `https://gymnasium.farama.org`

1. **Cliff Walking (`CliffWalking-v0`):** A deterministic 4×12 grid with start at $[3, 0]$ and goal at $[3, 11]$. A "cliff" spans $[3, 1]$–$[3, 10]$; stepping into it yields -100 and teleports the agent back to start. Each regular step yields -1, and the episode terminates upon reaching the goal. The observation space has 48 reachable states, and the action space has 4 discrete moves.

2. **Taxi (`Taxi-v3`):** A 5×5 grid in which a taxi must pick up a passenger at one of four fixed locations and deliver them to a specified destination. The observation space has size $|\mathcal{S}| = 500$ (25 taxi positions $\times$ 5 passenger locations $\times$ 4 destinations), and the action space $|\mathcal{A}| = 6$ (move south, north, east, west; pick up; drop off). Each step yields -1; illegal pick-up/drop-off yields -10; successful drop-off yields $+20$. Episodes end upon successful passenger delivery.

4.2 Implementation Details

All algorithms were implemented in Python 3.9 with Gymnasium 1.2.0 and NumPy 1.23, ensuring a pure tabular setting.

4.3 Experimental Protocol

All experiments employed a training budget of 100 000 independent episodes per environment. Each episode in Cliff Walking and Taxi was executed until the agent reached the goal state or a 700-step time limit was reached in the Cliff Walking environment and a 1500-step limit in the Taxi environment, such that the cumulative negative rewards model "suffering" that the agent minimizes. Rollback counts as a step even when no state change occurs. A fixed sequence of random seeds was applied systematically across all episodes and agents to ensure each algorithm experienced identical stochastic conditions. Statistical information-including episodic returns, rollback counts, and convergence metrics-was recorded for all 100 000 episodes and aggregated into CSV files for comprehensive post-hoc analysis.

4.4 Scope Justification

Many recent advances in reinforcement learning target high-dimensional or continuous control tasks; our study deliberately focuses on tabular environments to rigorously evaluate the proposed reversibility framework. Tabular benchmarks such as `CliffWalking-v0` and `Taxi-v3` allow us to isolate the effects of our empirical reversibility estimator and U-turn rollback mechanism without the confounding complexities introduced by function approximation or representation learning. By removing factors like neural network training dynamics and policy-gradient variance, we can precisely quantify how reversibility influences both safety (e.g., reduction in catastrophic transitions) and performance (e.g., steady improvement in cumulative return). Moreover, the deterministic nature of tabular implementations ensures complete reproducibility: every buffer update, estimator statistic, and rollback decision can be logged and inspected in full.

5 Experimental Evaluation and Results

In this subsection, we evaluate the impact of integrating reversibility and rollback into the Q-learning framework, focusing on mean performance, safety outcomes, and variance control in both `CliffWalking-v0` and `Taxi-v3` environments (Table 1). All reported statistics are computed over 100 000 episodes per agent configuration; 95%

confidence intervals are reported as $\bar{x} \pm 1.96\,\sigma/\sqrt{N}$. Rollback counts as a step for reward accounting in the next sub-sections.

Table 1. Comparison of performance and variance metrics between vanilla Q-learning and the reversibility-augmented agent.

Domain	Metric	Vanilla Q mean (CI)	σ_{van}	Rollback Precedence Q mean (CI)	σ_{mod}	Δ mean	%Δ mean	$\Delta\sigma$	%$\Delta\sigma$
CliffWalking-v0	Total Reward	$-399.77\,[-403.26, -396.27]$	563.78	$-179.81\,[-180.81, -178.81]$	160.97	$+219.96$	$+55.0\%$	-402.81	-71.4%
	Steps / episode	$181.06\,[180.08, 182.03]$	157.32	$182.89\,[181.85, 183.92]$	167.02	$+1.83$	$+1.01\%$	$+9.70$	$+6.2\%$
	Falls / episode	$2.20920\,[2.18351, 2.23489]$	4.14	$0.00370\,[0.00325, 0.00416]$	0.07	-2.2055	-99.8%	-4.07	-98.2%
	Rollbacks / episode	—	—	$3.4385\,[3.3927, 3.4843]$	7.39	$+3.4385$	n/a	$+7.39$	n/a
Taxi-v3	Total Reward	$-1652.93\,[-1656.98, -1648.88]$	652.74	$-567.09\,[-568.75, -565.44]$	267.00	$+1085.84$	$+65.7\%$	-385.74	-59.1%
	Steps / episode	$681.85\,[680.11, 683.60]$	281.22	$698.65\,[696.74, 700.56]$	308.49	$+16.80$	$+2.46\%$	$+27.27$	$+9.7\%$
	Illegal Drops / episode	$110.21690\,[109.95840, 110.47540]$	41.70	$0.06940\,[0.06764, 0.07116]$	0.28	-110.1475	-99.9%	-41.42	-99.3%
	Deliveries / episode	$0.99410\,[0.99362, 0.99458]$	0.077	$0.98500\,[0.98425, 0.98575]$	0.121	-0.00910	-0.92%	$+0.0450$	$+58.1\%$
	Rollbacks / episode	—	—	$111.5006\,[111.2280, 111.7732]$	43.98	$+111.5006$	n/a	$+43.98$	n/a

5.1 Performance and Safety in `CliffWalking-v0`

1. **Mean Episode Return:** The standard Q-learning agent attains an average return of -399.77 ($\sigma = 563.78$), whereas the reversibility-augmented agent achieves -179.81 ($\sigma = 160.97$), yielding a $+55.0\%$ reduction in penalty ($\Delta = +219.96$). This indicates that penalizing low-reversibility transitions and undoing unsafe moves steers the policy away from cliff-edge states.

2. **Catastrophic Falls:** Under vanilla Q-learning, the agent falls off the cliff 2.20920 times per episode ($\sigma = 4.14$). Introducing rollbacks reduces falls to 0.00370 per episode ($\sigma = 0.07$)-a -99.8% change. The rollback mechanism thus intercepts essentially all cliff transgressions before terminal penalty.

3. **Trajectory Efficiency:** Despite averaging 3.4385 corrective rollbacks per episode ($\sigma = 7.39$), the augmented agent's trajectories change from 181.06 steps ($\sigma = 157.32$) to 182.89 steps ($\sigma = 167.02$), a $+1.01\%$ shift. In this domain, safety comes at negligible path-length cost.

4. **Variance Control:** Variability in safety-critical quantities contracts sharply: return standard deviation drops by 71.4% ($563.78 \to 160.97$) and falls variance by 98.2% ($4.14 \to 0.07$). Path-length variability rises modestly ($157.32 \to 167.02$), consistent with occasional rollback-induced detours while preserving robust safety. This variance reduction effect is consistent with stabilization approaches such as Averaged-DQN [27], though our rollback mechanism achieves stability through explicit corrective interventions rather than ensemble averaging.

5.2 Performance and Safety in `Taxi-v3`

1. **Mean Episode Return:** Vanilla Q-learning yields -1652.93 ($\sigma = 652.74$), whereas the rollback-equipped agent reaches -567.09 ($\sigma = 267.00$), a $+65.7\%$ improvement ($\Delta = +1085.84$). Preventing illegal transitions before they accrue penalties recovers the bulk of negative reward.

2. **Illegal Action Suppression:** The frequency of illegal actions plunges from 110.21690 per episode ($\sigma = 41.70$) to 0.06940 ($\sigma = 0.28$)-a -99.9% change. The agent executes an average of 111.5006 rollbacks ($\sigma = 43.98$) per episode,

effectively catching nearly every invalid transition and avoiding the associated -10 penalties and wasted navigation.

3. **Trajectory Length and Success Rate:** Corrective rollbacks extend trajectories modestly: steps per episode rise from 681.85 ($\sigma = 281.22$) to 698.65 ($\sigma = 308.49$), a +2.46% increase. Delivery success declines slightly from 0.99410 to 0.98500 ($\Delta = -0.00910$, -0.92%; σ: $0.077 \rightarrow 0.121$), reflecting a more conservative policy that avoids risky shortcuts.

4. **Variance Control:** Return variance shrinks by 59.1% ($652.74 \rightarrow 267.00$) and illegal-action standard deviation by 99.3% ($41.70 \rightarrow 0.28$). Step-count variability rises by 9.7% ($281.22 \rightarrow 308.49$), and delivery variability increases (from 0.077 to 0.121), attributable to episodic fluctuations in rollback frequency and success outcomes. Overall, safety-critical metrics become markedly more predictable while modestly increasing path-length dispersion. This predictability aligns with prior work on deep exploration methods such as Bootstrapped DQN [28], though our approach reduces dispersion by constraining unsafe transitions rather than by bootstrapped value-function sampling.

5.3 Parameter Analysis

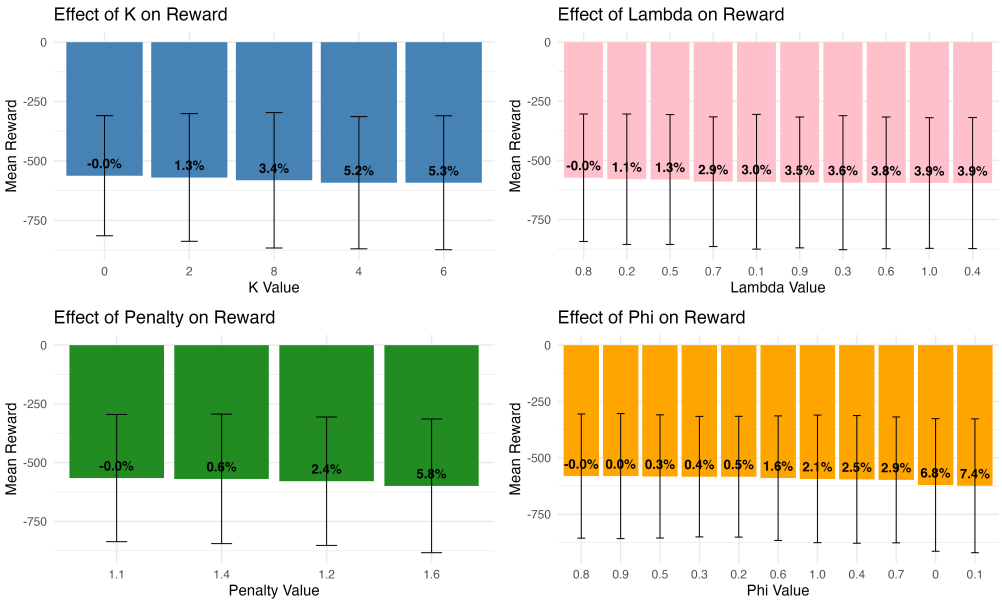

Fig 2. Parameter sensitivity analysis in `Taxi-v3`.

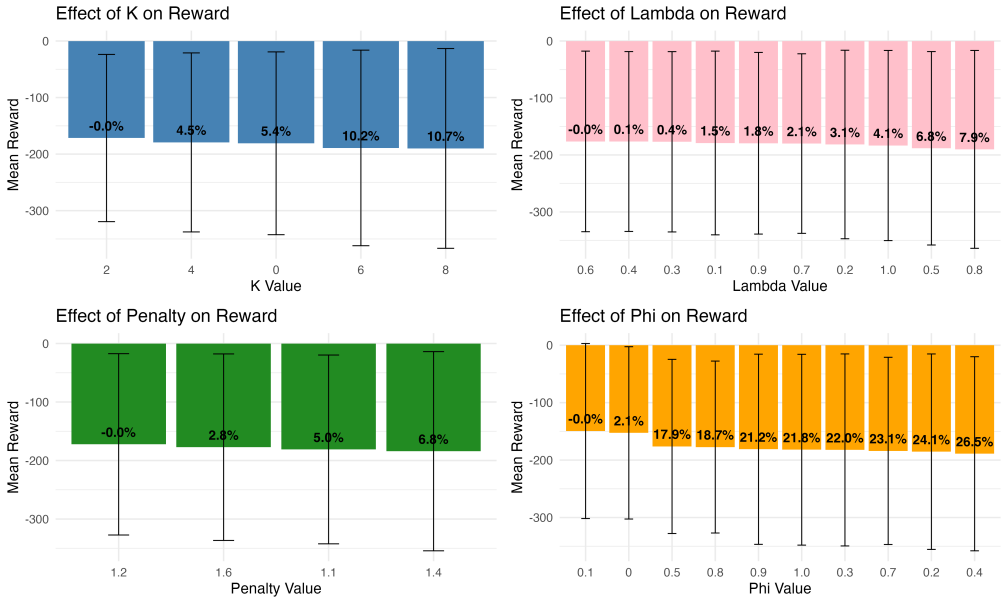

Fig 3. Parameter sensitivity analysis in `CliffWalking-v0`.

We now examine the sensitivity of the reversible learning framework to its four main parameters: horizon length (K), precedence learning rate (λ), penalty magnitude, and the initialization value of the reversibility estimator (Φ_0). For both domains, the first bar (Fig 2, Fig 3) in each sweep corresponds to the empirically optimal value, which we interpret before discussing degradation under alternative settings.

Horizon (K). In `CliffWalking-v0`, the optimal value is $K = 2$. This aligns with the environment's local grid dynamics, where safe reversals are typically only one or two steps away. Shorter windows (e.g., $K = 1$) miss legitimate reversals and cause excessive rollbacks, while longer horizons (e.g., $K = 4, 6$) dilute the local signal and mistakenly treat cliff-edge detours as reversible. Thus, reversibility in CliffWalking is predominantly local, and $K = 2$ best captures the true return structure.

By contrast, `Taxi-v3` exhibits an optimum at $K = 0$. Reversibility here is immediate: illegal pick-ups and drop-offs reveal themselves instantly, and grid navigation is inherently safe. Any extension of the horizon introduces noise from loops in the taxi's movement, delaying rollback corrections. Performance degrades monotonically with larger K, with $K = 6$–8 being the weakest. This contrast illustrates that CliffWalking benefits from short local windows, while Taxi rewards purely instantaneous checks.

Precedence Learning Rate (λ). For `CliffWalking-v0`, the optimal setting is $\lambda = 0.6$, with 0.4 and 0.3 also performing strongly. Smaller values (e.g., 0.1) underfit reversibility signals, while larger extremes destabilize updates. This indicates that CliffWalking favors a relatively fast but stable reversibility learner.

In `Taxi-v3`, the optimal value is $\lambda = 0.8$, with weaker but still viable performance at 0.2–0.5. Too-slow rates again lag behind environmental evidence, while non-optimal values like 1.0 or 0.4 inject noise. Because Taxi features many repeated sub-trajectories, rapid updates to Φ are necessary to keep rollback triggers aligned with the current episode dynamics.

Penalty Magnitude. In `CliffWalking-v0`, the best result is at penalty = 1.2, followed closely by 1.6 and 1.1. The weakest was 1.4, with a 6.8% performance drop

relative to the optimum. This suggests that penalties clustered around 1.1–1.6 work well, but tuning is important: too low under-corrects, while poorly calibrated values (like 1.4) disrupt efficiency.

In `Taxi-v3`, the optimum is 1.1, with 1.4 and 1.2 still serviceable. However, 1.6 produced the weakest performance, over-constraining exploration. Since Taxi already imposes large native penalties (-10 for illegal moves), a lighter reversibility penalty is sufficient; higher values introduce unnecessary rollback frequency.

Initialization Value (Φ_0). The `CliffWalking-v0` domain is best served by $\Phi_0 = 0.0$ or 0.1. This pessimistic prior reflects the environment's high asymmetry between safe moves and catastrophic cliff falls. By assuming most transitions are irreversible until proven otherwise, the agent leverages rollback early and avoids premature overconfidence near the cliff. More optimistic initializations (e.g., 0.5–1.0) performed substantially worse, as they caused misjudgments of danger and frequent falls.

The opposite holds in `Taxi-v3`, where the optimal initialization lies around $\Phi_0 = 0.8$–0.9. Because Taxi contains many inherently reversible transitions (safe grid navigation), an optimistic prior reduces unnecessary rollbacks and penalties on benign moves. Pessimistic values (e.g., $\Phi_0 = 0.0$–0.1) misclassify ordinary movements as irreversible, inflating rollbacks and hurting efficiency.

Summary. Taken together, the parameter sweeps reveal environment-specific sensitivities. `CliffWalking-v0` rewards a short local horizon ($K = 2$), a moderately fast precedence learner ($\lambda = 0.6$), a carefully tuned penalty near 1.2, and a pessimistic initialization ($\Phi_0 = 0.0$–0.1) reflecting its hazardous structure. `Taxi-v3`, in contrast, favors hyper-local checks ($K = 0$), a fast learner ($\lambda = 0.8$), a lighter penalty (1.1), and an optimistic prior ($\Phi_0 = 0.8$–0.9). These contrasts underscore that reversibility-aware RL is not governed by a single "best" hyperparameter profile but must adapt its biases: CliffWalking demands caution and pessimism near irreversible cliffs, while Taxi thrives with optimism, immediacy, and lighter corrective signals.

5.4 Parameter Sensitivity

The effectiveness of the reversibility + rollback framework critically depends on two key parameters:

Q-Table Initialization Value (Q_0)

In the modified algorithm, initializing all Q-values to zero biases the penalty-and-rollback criterion: zero Q-values can cause the rollback condition to misfire, leading to suboptimal or inconsistent rollbacks. We therefore initialize for both environments

$$Q_0 = -1.$$

This was the optimal initialization value given the reward structure in both domains.

Penalty Threshold (T)

The rollback criterion fires when the reversibility-penalized TD target falls below

$$T \cdot Q(s, a).$$

If T is too high, legitimate exploratory moves are rolled back excessively, over-constraining the policy; if T is too low, unsafe transitions may slip through uncorrected. We select T empirically based on the domain's reward scale (e.g., $T = 3$ for both `CliffWalking-v0` and `Taxi-v3`) to balance safety intervention against necessary exploration.

An incorrect threshold choice can either

(a) suppress learning by over-rolling back, or

(b) fail to prevent catastrophic events,

resulting in skewed performance metrics and increased variance.

5.5 Ablation Study

We disentangle the effects of three components-*rollback*, *threshold-based penalization*, and *precedence (Φ) penalties*-across `CliffWalking-v0` and `Taxi-v3`. Agent configurations and hyperparameters are listed in Table 2; outcome metrics are reported in Tables 3, 4, and the attribution-style summary in Table 5.

Table 2. Parameter matrix for agents in `CliffWalking-v0` and `Taxi-v3`.

Env	Agent	α	γ	ϵ	*q_table_init*	K	α_ϕ	λ_{prec}	ϕ_{init}	*threshold*	*penalty*
`CliffWalking-v0`	Baseline (QL)	0.1	0.99	0.1	0.0	–	–	–	–	–	–
	RollbackOnly	0.1	0.99	0.1	−1.0	–	–	–	–	3	–
	ThresholdPeAgent	0.1	0.99	0.1	−1.0	–	–	–	–	3	1.1
	Roll_Threshold	0.1	0.99	0.1	−1.0	–	–	–	–	3	1.1
	PrecedenceOnly	0.1	0.99	0.1	−1.0	2	0.01	0.6	0.1	–	–
	Precedence_R	0.1	0.99	0.1	−1.0	2	0.01	0.6	0.1	3	–
	Precedence_Th	0.1	0.99	0.1	−1.0	2	0.01	0.6	0.1	3	1.1
	FullModel	0.1	0.99	0.1	−1.0	2	0.01	0.6	0.1	3	1.1
`Taxi-v3`	Baseline (QL)	0.1	0.99	0.1	0.0	–	–	–	–	–	–
	RollbackOnly	0.1	0.99	0.1	−1.0	–	–	–	–	3	–
	ThresholdPeAgent	0.1	0.99	0.1	−1.0	–	–	–	–	3	1.1
	Roll_Threshold	0.1	0.99	0.1	−1.0	–	–	–	–	3	1.1
	PrecedenceOnly	0.1	0.99	0.1	−1.0	2	0.01	0.8	0.8	–	–
	Precedence_R	0.1	0.99	0.1	−1.0	2	0.01	0.8	0.8	3	–
	Precedence_Th	0.1	0.99	0.1	−1.0	2	0.01	0.8	0.8	3	1.1
	FullModel	0.1	0.99	0.1	−1.0	2	0.01	0.8	0.8	3	1.1

Table 3. Ablation results on CliffWalking-v0. Rewards are averaged with standard deviation. Δ values are relative improvements over the baseline.

Agent	Reward	Δ Reward	Δ%	Failures	Δ Fail%	Rollbacks
Roll_Threshold	-174.4 ± 151.4	+225.3	+56.4%	0.004	+99.8%	2.3
RollbackOnly	-174.9 ± 152.3	+224.8	+56.2%	0.004	+99.8%	2.4
FullModel	-179.8 ± 161.0	+220.0	+55.0%	0.004	+99.8%	3.4
Precedence_R	-181.5 ± 162.8	+218.3	+54.6%	0.004	+99.8%	3.5
ThresholdPeAgent	-398.2 ± 566.1	+1.6	+0.4%	2.174	+1.6%	n/a
Baseline	-399.8 ± 563.8	+0.0	+0.0%	2.209	n/a	n/a
Precedence_Th	-424.1 ± 605.4	−24.3	−6.1%	2.354	−6.6%	n/a
PrecedenceOnly	-427.5 ± 609.3	−27.8	−6.9%	2.378	−7.7%	n/a

Table 4. Ablation results on Taxi-v3. Rewards are averaged with standard deviation. Δ values are relative improvements over the baseline.

Agent	Reward	Δ Reward	Δ%	Failures	Δ Fail%	Rollbacks
RollbackOnly	-551.8 ± 241.7	+1101.2	+66.6%	0.033	+100.0%	110.3
Roll_Threshold	-552.0 ± 241.0	+1101.0	+66.6%	0.063	+99.9%	110.2
FullModel	-567.1 ± 267.0	+1085.8	+65.7%	0.069	+99.9%	111.5
Precedence_R	-567.7 ± 266.0	+1085.2	+65.7%	0.017	+100.0%	111.7
Baseline	-1652.9 ± 652.7	+0.0	+0.0%	110.217	n/a	n/a
ThresholdPeAgent	-1654.2 ± 654.1	−1.2	−0.1%	110.269	−0.0%	n/a
Precedence_Th	-1683.2 ± 699.7	−30.2	−1.8%	111.632	−1.3%	n/a
PrecedenceOnly	-1686.1 ± 702.1	−33.1	−2.0%	111.805	−1.4%	n/a

Table 5. Component contribution analysis for CliffWalking-v0 and Taxi-v3. Baseline refers to vanilla Q-learning without rollback, threshold, or Φ-penalty.

Environment	Configuration	Reward Improvement	Share of Full Model	Failure Reduction	Rollbacks / Episode
CliffWalking-v0	Baseline (Q-Learning)	−399.8 reward, 2.209 fails	—	—	n/a
	Rollback Only	+224.8 (+56.2%)	102.2%	+2.205 (+99.8%)	2.4
	Precedence Only	−27.8 (−6.9%)	−12.6%	−0.169 (−7.7%)	n/a
	Full Model (All comps.)	+220.0 (+55.0%)	100.0%	+2.206 (+99.8%)	3.4
Taxi-v3	Baseline (Q-Learning)	−1652.9 reward, 110.217 fails	—	—	n/a
	Rollback Only	+1101.2 (+66.6%)	101.4%	+110.184 (+100.0%)	110.3
	Precedence Only	−33.1 (−2.0%)	−3.1%	−1.588 (−1.4%)	n/a
	Full Model (All comps.)	+1085.8 (+65.7%)	100.0%	+110.147 (+99.9%)	111.5

We ablate three components-explicit rollback, threshold-based scaling, and precedence (Φ) penalties-across `CliffWalking-v0` and `Taxi-v3` under identical tabular Q-learning settings and training budgets. Metrics are mean return, failure rate (falls or illegal actions), rollback frequency, and dispersion (SD), computed over 100,000 episodes per agent.

Rollback is the dominant driver of both safety and performance: RollbackOnly and Roll_Threshold recover essentially all of the full model's reward improvement while virtually eliminating failures ($\geq$99.8%). By contrast, PrecedenceOnly underperforms vanilla Q-learning in both domains, indicating that Φ-penalties alone misguide updates when self-transitions and resets are frequent. Thresholding is secondary: it contributes little on its own and adds value primarily when paired with rollback, with gains that depend on the environment.

In `CliffWalking-v0`, Roll_Threshold achieves the best mean return (−174.4) with RollbackOnly a close second (−174.9). Both exceed FullModel (−179.8) while maintaining the same near-zero failure rate. Notably, Roll_Threshold attains the top return with fewer rollbacks per episode (2.3) than FullModel (3.4), suggesting that once catastrophic moves are suppressed, the threshold prunes unnecessary reversions and slightly improves path efficiency.

In `Taxi-v3`, RollbackOnly is strongest (−551.8), narrowly ahead of Roll_Threshold and clearly ahead of FullModel. Adding Φ reduces returns without measurable safety gains (failures are already ≈ 0 under rollback). Thresholding does not meaningfully change rollback usage (110.2 vs 110.3), indicating limited leverage in navigation-dominated regimes where frequent, benign reversions are intrinsic to task structure.

Across both tasks, rollback variants markedly compress reward dispersion (e.g., Cliff

SD ≈151–162 vs Baseline ≈564; Taxi SD ≈241–267 vs Baseline ≈653), consistent with smoother learning trajectories. This variance collapse, coupled with order-of-magnitude failure reductions, supports the view that reversible corrections prevent catastrophic updates from propagating.

Plotting return against rollbacks per episode yields a Pareto-like frontier dominated by rollback agents. In `CliffWalking-v0`, Roll_Threshold occupies a favorable corner (better return and fewer rollbacks than FullModel). In `Taxi-v3`, the frontier is essentially flat between RollbackOnly and Roll_Threshold, implying that thresholding adds little efficiency once rollback usage saturates.

Mechanistically, rollback acts as a local safety filter that caps downside by immediately reversing low-quality transitions before value errors spread-akin to a risk-sensitive control at the transition level. Thresholding regulates the rollback budget, helping in cliff-like domains where failures are sparse but costly. Φ-penalties pressure the agent away from high-precedence (hard-to-undo) regions, but in environments with many benign self-transitions (e.g., `Taxi-v3`) this shaping conflates necessary loops with hazards, degrading policy quality unless guarded by rollback.

Practical guidance: use explicit rollback as the default safety primitive; add thresholding to trim extraneous reversions in cliff-like tasks; apply Φ-penalties sparingly and only alongside rollback, tuning them with awareness of self-transition prevalence. This recipe preserves the safety guarantee, captures most of the performance lift, and controls variance.

6 Discussion

The results show that embedding reversibility into reinforcement learning improves both safety and performance across environments. In `CliffWalking-v0`, reversibility-aware agents reduced catastrophic failures by $\geq 99.8\%$ while substantially improving cumulative returns and compressing variance. In `Taxi-v3`, selective rollback suppressed illegal actions by $\geq 99.9\%$, transforming persistent penalties into recoverable states. Together, these outcomes indicate that reversibility not only mitigates overestimation-induced errors but also acts as a variance-control mechanism that stabilizes learning in safety-critical domains.

Ablations isolate *rollback* as the primary driver of these gains. Rollback-only and rollback+threshold agents recover essentially all of the full model's return improvements while preserving the near-elimination of failures. By contrast, *precedence* (Φ) penalties without rollback underperform vanilla Q-learning in both tasks; with rollback, their contribution is *environment-dependent*: in CliffWalking they are at best marginally helpful (and sometimes neutral) once failures are already suppressed, whereas in Taxi they tend to degrade returns due to frequent benign self-transitions being misclassified as undesirable. Hence, hard interventions (explicit undo) dominate soft shaping; and any Φ-based shaping should be used sparingly and only alongside rollback.

Parameter sensitivity reinforces this environment-specific picture. CliffWalking benefits from pessimistic priors on reversibility, short horizons, and moderate penalties-consistent with highly asymmetric costs (safe moves vs. cliff falls). Taxi favors optimistic priors, immediate rollbacks, and lighter penalties, reflecting its abundance of inherently reversible transitions and the large native penalty already attached to invalid actions. These contrasts confirm that reversibility-aware RL is not "one-size-fits-all": it should encode environment-specific biases to trade off caution and efficiency.

Our results are limited to tabular Gym/Gymnasium toy-text domains (e.g., `CliffWalking-v0`, `Taxi-v3`), selected for transparency and controllability rather than representational richness. Consequently, conclusions about safety and variance reduction may not carry over without modification to function-approximation settings

or high-dimensional continuous control. The rollback operator further assumes access to a safe *previous-state* primitive (or an equivalent reset/checkpoint facility). While this is realistic in simulated grids, it can be non-trivial in real systems; even when available. Finally, effectiveness is sensitive to environment-aware hyperparameter selection—horizon K, threshold T, penalty scale λ, and Φ initialization (Φ_0) which requires tuning prior to deployment or extension to deep function approximation.

Future research should focus on experimenting with the integration of Rollback in function approximation settings and expanding the experimental domains for precedence estimation to narrow down the use cases for precedence estimation usability in terms of performance. Furthermore, this work can be considered a foundation for behavior modeling, as precedence and rollback can be utilized in encoding agent behavior profiles as optimistic, pessimistic, high or low tolerance to risk, and patience level modeling, which provides foundations for conditions in the human decision-making process.

7 Conclusion

We introduce a reversible reinforcement learning framework that couples an empirical reversibility estimator with an explicit rollback operator and, across two benchmark environments, delivers (1) substantial safety gains—over 99% fewer catastrophic failures and illegal actions, (2) improved performance—roughly 55–66% higher cumulative reward than vanilla Q-learning, (3) variance control—markedly lower dispersion in both reward and safety metrics, and (4) environment-specific adaptability—distinct optimal parameterizations for hazardous versus benign domains. Ablations identify rollback as the critical mechanism; thresholding further improves rollback efficiency in cliff-like tasks, while precedence estimation is supportive, strongly context dependent, and harmful if applied without rollback. Overall, reversibility emerges as a practical, powerful organizing principle for safety-sensitive RL. Future work should extend the framework to deep function approximation, develop adaptive hyperparameter tuning across environments, and investigate real-world analogues of rollback in robotics and decision-support systems. By operationalizing the ability to "undo" mistakes, reversibility-aware RL advances the design of safe, robust, and trustworthy autonomous agents and also can be see as foundation for behavior modeling in decision making agents.

References

1. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed. MIT Press; 2018.

2. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. arXiv preprint arXiv:160606565. 2016. Available from: `https://arxiv.org/abs/1606.06565`.

3. Thrun S, Schwartz A. Issues in using function approximation for reinforcement learning. In: Mozer M, Smolensky P, Touretzky D, Elman J, Weigend A, editors. Connectionist Models: Proceedings of the 1993 Summer School. Lawrence Erlbaum; 1993. .

4. van Hasselt H, Guez A, Silver D. Deep Reinforcement Learning with Double Q-learning. arXiv preprint arXiv:150906461. 2015. Available from: `https://arxiv.org/abs/1509.06461`.

5. Garcia J, Fernández F. A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research. 2015;16:1437-80.

6. Guerrier M, Fouad H, Beltrame G. Learning Control Barrier Functions and Their Application in Reinforcement Learning: A Survey. 2024. arXiv:2404.16879.

7. Zhao W, He T, Chen R, Wei T, Liu C. State-wise Safe Reinforcement Learning: A Survey. In: Proceedings of IJCAI 2023, Survey Track; 2023. p. 6814-22.

8. Wachi A, Shen X, Sui Y. A Survey of Constraint Formulations in Safe Reinforcement Learning. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI); 2024. .

9. Kushwaha A, Ravish K, Lamba P, Kumar P. A Survey of Safe Reinforcement Learning and Constrained MDPs: A Technical Survey on Single-Agent and Multi-Agent Safety. 2025. arXiv:2505.17342.

10. Gu S, Yang L, Du Y, Chen G, Walter F, Knoll A. A Review of Safe Reinforcement Learning: Methods, Theories, and Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024;46(12):11216-35. doi:10.1109/TPAMI.2024.3457538.

11. Lan Q, Pan Y, Fyshe A, White M. Maxmin Q-learning: Controlling the Estimation Bias of Q-learning. In: International Conference on Learning Representations (ICLR); 2020. Available from: `https://arxiv.org/abs/2002.06487`.

12. Grinsztajn N, Ferret J, Pietquin O, Preux P, Geist M. There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning. arXiv preprint arXiv:210604480. 2021. Available from: `https://arxiv.org/abs/2106.04480`.

13. Watkins CJCH. Learning from Delayed Rewards [Ph.D. thesis]. Cambridge, England: University of Cambridge; 1989.

14. van Hasselt H. Double Q-learning. In: Advances in Neural Information Processing Systems; 2010. .

15. Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. arXiv preprint arXiv:180209477. 2018. Available from: `https://arxiv.org/abs/1802.09477`.

16. Moldovan TM, Abbeel P. Safe exploration in Markov decision processes. In: International Conference on Machine Learning; 2012. p. 1711-8.

17. Wachi A, Hashimoto W, Shen X, Hashimoto K. Safe Exploration in Reinforcement Learning: A Generalized Formulation and Algorithms. In: Advances in Neural Information Processing Systems (NeurIPS); 2023. .

18. As Y, Sukhija B, Treven L, Sferrazza C, Coros S, Krause A. ActSafe: Active Exploration with Safety Constraints for Reinforcement Learning. In: International Conference on Learning Representations (ICLR); 2025. .

19. Wang Y, Zhu H. Safe Exploration in Reinforcement Learning by Reachability Analysis over Learned Models. In: Computer Aided Verification (CAV); 2024. .

20. Gu S, Sel B, Ding Y, Wang L, Lin Q, Jin M, et al. Balance Reward and Safety Optimization for Safe Reinforcement Learning: A Perspective of Gradient Manipulation. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2025. .

21. Achiam J, Held D, Tamar A, Abbeel P. Constrained Policy Optimization. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017. p. 22-31.

22. Ray A, Achiam J, Amodei D. Benchmarking Safe Exploration in Deep Reinforcement Learning. In: Proceedings of the 2019 Safe Machine Learning Workshop at ICLR; 2019. ArXiv preprint arXiv:1910.01708.

23. Eysenbach B, Gupta A, Ibarz J, Levine S. Diversity is all you need: Learning skills without a reward function. In: International Conference on Learning Representations; 2019. .

24. Harutyunyan A, Bellemare MG, Stepleton T, Munos R. Q(λ) with off-policy corrections. In: Advances in Neural Information Processing Systems; 2016. p. 6975-83.

25. Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, et al. OpenAI Gym. arXiv preprint arXiv:160601540. 2016. Available from: `https://arxiv.org/abs/1606.01540`.

26. Terry JK, Black B, Grammel N, Jayakumar M, Santos L, Sulivan R, et al. PettingZoo: Gym for multi-agent reinforcement learning. In: Advances in Neural Information Processing Systems; 2020. .

27. Anschel O, Baram N, Shimkin N. Averaged-DQN: Variance reduction and stabilization in deep reinforcement learning. In: International Conference on Machine Learning; 2017. p. 176-85.

28. Osband I, Blundell C, Pritzel A, Van Roy B. Deep exploration via bootstrapped DQN. In: Advances in Neural Information Processing Systems; 2016. p. 4026-34.