SAFE POLICY CORRECTION VIA TARGETED UNLEARNING IN REINFORCEMENT LEARNING

TAMÁS TAKÁCS, LÁSZLÓ GULYÁS

tamastheactual@inf.elte.hu
lgulyas@inf.elte.hu

A Three-Pillar Framework: Trajectory-Selective Forgetting, Zero-Shot Strategy Inversion, and Metaplasticity-Based Retain Protection

Bosch B.R.I.D.G.E. meetup

ELTE | IK INFORMATIKAI KAR

MESTERSÉGES INTELLIGENCIA TANSZÉK

BOSCH

1. INTRODUCTION/MOTIVATION

Reinforcement learning agents deployed in production systems develop undesired behaviours during operation that require correction. Traditional approaches discard the trained policy and restart from scratch, wasting computational resources. Machine Unlearning offers an alternative: selectively remove specific behaviours while preserving useful knowledge. Existing SOTA methods face critical limitations. *TrajDeleter* **[1]** achieves 94.8% forget effectiveness but operates only on offline RL with fixed trajectory buffers. *Reinforcement Unlearning* **[2]** addresses environment-level forgetting but lacks behaviour-level granularity. Neither handles online/on-policy scenarios where agents continue interacting with environments during deployment.

2. METHODS AND MATERIALS

We address selective behavior unlearning in on-policy RL. Given a trained policy π_θ parameterized by θ and a set of undesired behaviors $\mathcal{D}_f$ (unsafe trajectories, exploitative action sequences), we construct an updated policy $\pi_{\theta'}$ satisfying **forget effectiveness** ($E_{\tau\sim\mathbb{D}_f}[\pi_{\theta'}(\tau)] \ll E_{\tau\sim\mathbb{D}_f}[\pi_\theta(\tau)]$), **retain stability** ($E_{\tau\sim\mathbb{D}_r}[J(\pi_{\theta'})] \geq \alpha \cdot E_{\tau\sim\mathbb{D}_r}[J(\pi_\theta)]$) maintaining performance on desired behaviors $\mathcal{D}_r$ (e.g. cart velocity or position in CartPole), and **computational efficiency** (unlearning cost substantially lower than retraining). Our three-pillar framework provides complementary solutions for different deployment constraints.

3. DISCUSSION

Trajectory-Selective Forgetting decomposes the unlearning objective into three loss components. A forget loss applies gradient reversal, maximizing negative log-likelihood to decrease probability of toxic actions. A retain loss maintains performance on acceptable behaviours through standard policy gradient updates **Zero-Shot Strategy Inversion** addresses scenarios where explicit toxic trajectories are unavailable due to privacy constraints or when undesirable behaviours are specified abstractly. We perform *policy inversion* by optimizing states rather than parameters. Sample seed states $s_0 \sim \text{Uniform}(\Omega)$ from the observation space, then solve $s^* = \arg\min_{s\in\Omega}\left[-\log \pi_\theta\left(a_{\text{target}}|s\right) + \beta|s - s_0|^2\right]$ via gradient descent on the state space. **Metaplasticity-Based Retain Protection** prevents catastrophic forgetting through a dual mechanism. We compute parameter importance via Fisher Information approximation I_F^I which captures sensitivity to perturbations. Unlike EWC, which adds a Fisher-weighted quadratic penalty to the loss, our MRP approach uses Fisher-based binary masks that rescale gradients dynamically via metaplasticity-inspired damping.

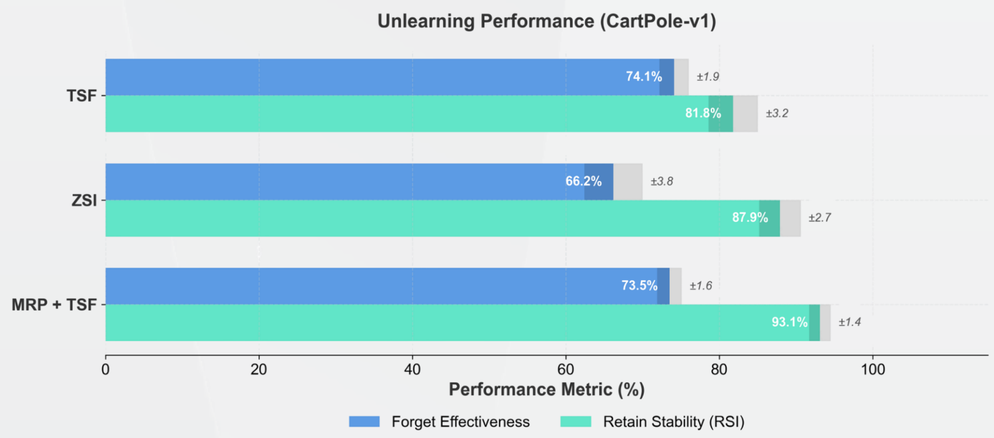


Figure 1: CartPole-v1 Benchmarks. TSF (top) is the most efficient solution with just 4.1% overhead. MRP+TSF (bottom) provides the best safety profile, retaining 93.1% of original performance. All methods achieve effective unlearning at a fraction (<7%) of retraining costs.

Metric	TSF	ZSI	MRP
Forget Effectiveness	**73.4 ± 2.1%**	65.8 ± 4.2%	72.9 ± 1.8%
Retention Stability (RSI)	82.1 ± 3.5%	88.3 ± 2.9%	**93.4 ± 1.5%**
Computational Overhead	**4.2 ± 0.8%**	6.7 ± 1.2%	5.1 ± 0.9%
Trajectory Requirements	Required	**Not required**	Required
Memory Complexity	$\mathcal{O}(\lvert\mathcal{D}_f\rvert)$	$\mathcal{O}(N_{\text{seeds}})$	$\mathcal{O}(\lvert\theta\rvert)$
Loss Function	$\alpha\mathcal{L}_f + \beta\mathcal{L}_r + \gamma\mathcal{L}_{\text{reg}}$	$\arg\min_s -\log\pi_\theta(a\lvert s)$	I_F mask + KL
Key Innovation	Explicit loss decomposition	Policy inversion without data	Parameter-level protection
Deployment Scenario	General-purpose unlearning	Privacy-constrained environments	Safety-critical systems

Table 1: A summary of the three proposed frameworks highlighting their distinct operational mechanisms and trade-offs. TSF leverages explicit loss decomposition for efficiency; ZSI utilizes zero-shot policy inversion for privacy; and MRP employs metaplasticity masks for safety-critical retention. Quantitative metrics (mean ± std. dev.) reflect mean performance on all three environments, based on 5 seeded runs each..

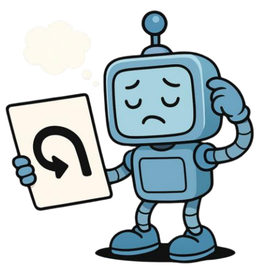
4. CONCLUSION/RESULTS

We validated our framework across three distinct Gymnasium environments with minor modifications, compatible with PPO: CartPole, LunarLander, and a MountainCar. The PPO implementation uses a shared MLP feature extractor with orthogonal initialization ($\sqrt{2}$) feeding into separate linear heads for the Actor and Critic. This setup shares parameters for feature learning but keeps policy and value estimation separate. We evaluate unlearning performance using **Forget Effectiveness** (derived from Area Under Forget Curve), **Retain Stability Index (RSI)**, and **Computational Overhead**.

Experimental results demonstrate that TSF achieves **73.4 ± 2.1%** Forget Effectiveness and **82.1 ± 3.5%** RSI at a minimal **4.2 ± 0.8%** computational overhead. In privacy-constrained scenarios, ZSI trades effectiveness (**65.8 ± 4.2%**) for enhanced RSI (**88.3 ± 2.9%**) with **6.7 ± 1.2%** overhead. For safety-critical applications, combining MRP with TSF maintains **72.9 ± 1.8%** effectiveness while boosting RSI to **93.4 ± 1.5%** at **5.1 ± 0.9%** overhead. This work proposes a novel framework architecture, currently at the proof-of-concept stage with validation only in controlled and simulated environments.

Our framework addresses three key limitations in existing work. First, it extends unlearning from offline RL to online/on-policy scenarios where agents continue environmental interaction. Second, **Zero-Shot Strategy Inversion** introduces an unlearning method without explicit trajectory examples, using gradient-based state optimization $\nabla_s \log \pi_\theta\left(a_{\text{toxic}}|s\right)$ to discover policy vulnerabilities. Third, explicit retain protection via Fisher Information masks and distillation provides tuneable safety guarantees, contrasting with implicit retention strategies. Our modular design allows independent or combined deployment: **TSF** for speed, **ZSI** for privacy, **MRP** for safety-critical systems.

5. OUTLOOK/REFERENCES



Our long-term directions include formal privacy analysis via **membership inference attacks** and **differential privacy guarantees**, multi-agent coordination where multiple policies require synchronized unlearning, and model-based integration combining our framework with world models for sample efficiency. Future work will explore **influence functions** for precise forgetting, parameter isolation and robust retention.

[1] Gong, C., Li, K., Yao, J., & Wang, T. (2024). TrajDeleter: Enabling Trajectory Forgetting in Offline Reinforcement Learning Agents. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2404.12530
[2] Ye, D., Zhu, T., Zhu, C., Wang, D., Gao, K., Shi, Z., … Xue, M. (2024). Reinforcement Unlearning. *arXiv [Cs.CR]*. Retrieved from http://arxiv.org/abs/2312.15910