Brittle Unlearning in On-Policy Reinforcement Learning

Cosmetic vs. structural forgetting under the relearn attack

Tamás Takács, László Gulyás

0009-0006-3027-7491, 0000-0002-6367-6695

{tamastheactual, lgulyas}@inf.elte.hu

ELTE Eötvös Loránd University, Faculty of Informatics, Department of Artificial Intelligence

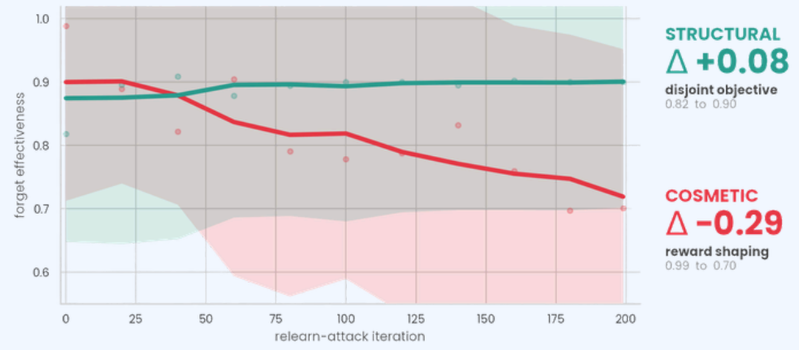
ELTE | IK INFORMATIKAI KAR

MESTERSÉGES INTELLIGENCIA TANSZÉK

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal

Abstract

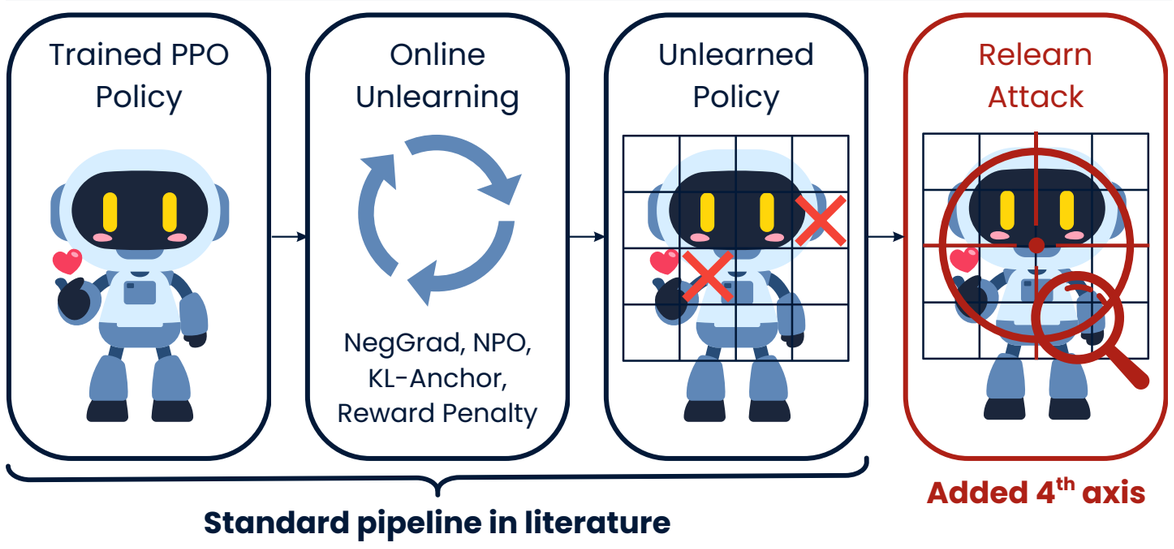
Unlearning recipes from supervised learning (NegGrad, NPO, KL-anchored variants, reward shaping, and Lagrangian PPO) are adapted to on-policy reinforcement learning and stress-tested under a continued-PPO relearn attack. Two findings emerge. First, every reward-shaping and gradient-ascent method produces a cosmetic forget that crashes when PPO resumes training. Second, a single change to the PPO update, zeroing the advantage on forget transitions, yields a structural forget that survives the attack. The effect holds across 3 environments, 7 methods, and 5 seeds, and depends on whether the forget region was reward-aligned in the original task.



Reward-shaping and gradient-ascent RL unlearning **produce cosmetic forgets** that crash under continued PPO. PPO-advantage-disjoint unlearning is **structural and survives the attack. The disjoint update removes forget transitions from PPO's objective entirely.** At every rollout step the scenario predicate produces a boolean indicator $\text{is\_forget}(s_t) = 1_{\mathcal{F}}(s_t)$, equal to 1 when the state lies in the forget region and 0 otherwise. The **advantage is gated by its complement,** $\tilde{A}_t = (1 - \text{is\_forget}(s_t)) \cdot \hat{A}_t$, so the **surrogate-loss contribution of any forget transition is identically zero,** and every retain transition keeps its full advantage and drives the PPO gradient. Retain is therefore not a separately labelled class. It is everything the predicate did not flag.

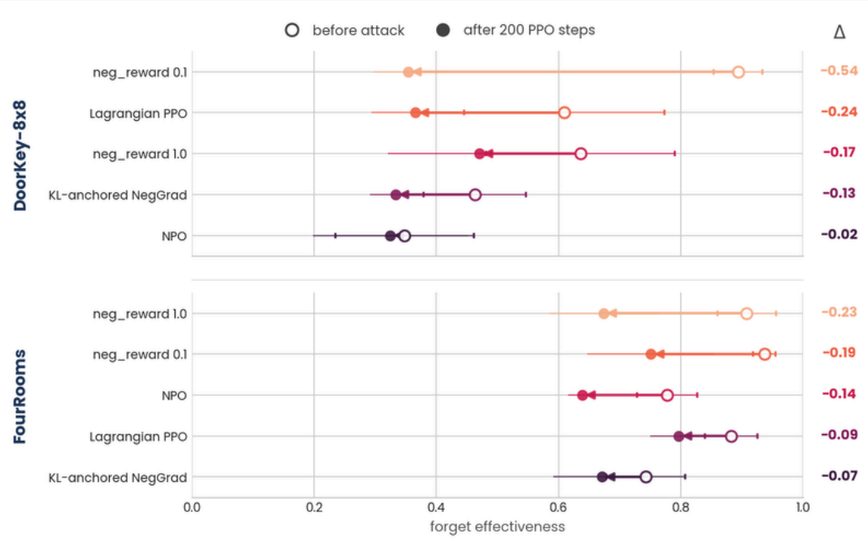
1. We are the **first to port classic unlearning methods into on-policy online RL,** and every one collapses under continued training.
2. Across 3 environments and 7 methods, **every reward-shaping and gradient-ascent unlearning crashes** within 200 PPO update steps.
3. By **cutting PPO's reinforcement signal on the forget transitions**, the policy must find an alternative path, and the **unlearning survives the attack.**

Data-free On-policy RL Unlearning Needs a Relearn-Resistance Axis



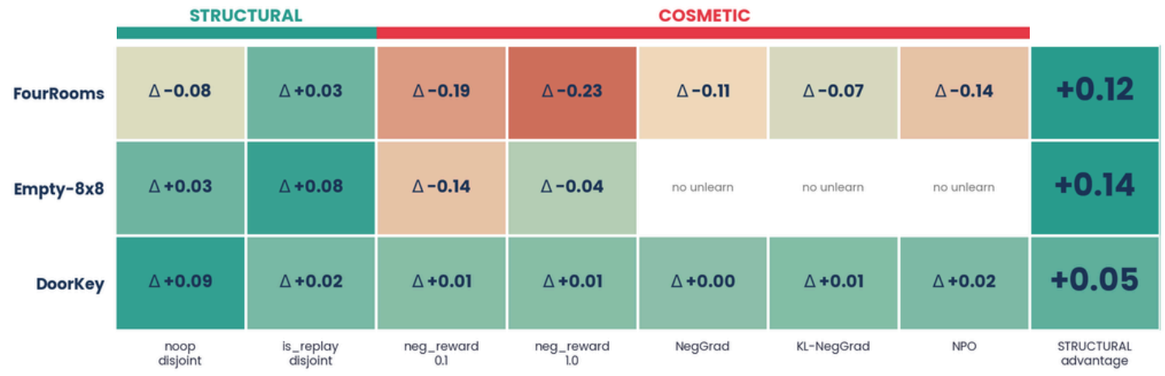
We adapt unlearning recipes from supervised learning to the on-policy RL setting using only the trained policy as a sampler. We then measure whether the forgetting survives a continued-training attack.

Penalty Methods are Cosmetic



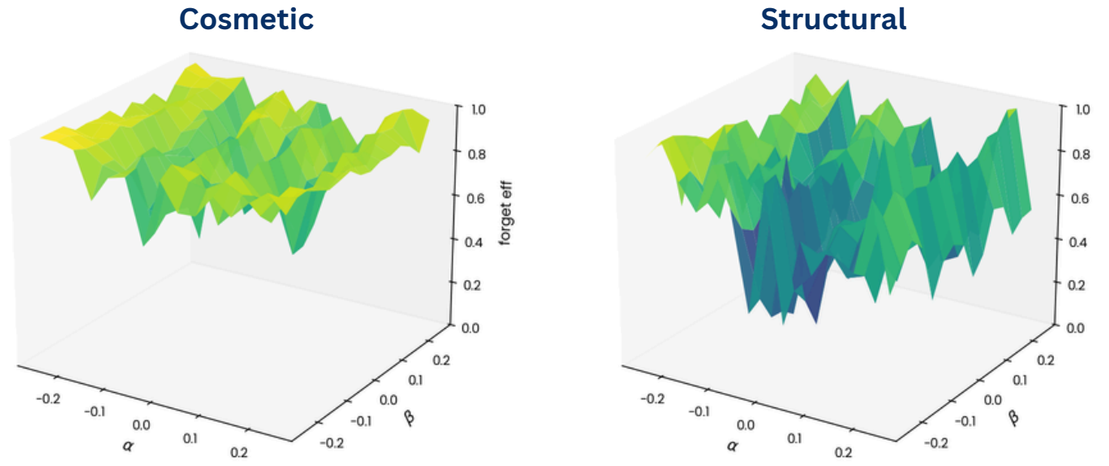
Five published unlearning methods all crash by 7–54 percentage points within 200 PPO update steps on both MiniGrid envs. The "before" score was suppression on top of an unchanged policy.

Cosmetic Unlearning is the Rule, not the Exception



Forget-effectiveness change (Δ) after a 200-iter PPO relearning attack on each unlearned policy, averaged over 5 seeds. Structural methods constrain the policy through a disjoint forget/retain objective, while cosmetic methods penalize the forget region without that separation. Structural cells stay near zero or positive across all envs, cosmetic cells lose forget under continued training on FourRooms and Empty-8x8, and three cosmetic methods fail to unlearn outright on Empty-8x8. The family-mean gap is positive in every environment, smallest on DoorKey where every method clusters near its natural forget floor.

One Architectural Change Makes Unlearning Structural



Each surface is forget effectiveness on a 2D filter-normalized random perturbation of the unlearned weights. The cosmetic policy sits on a broad plateau that penalty suppression manufactures, with a cliff along PPO's actual gradient. The structural policy lives at the env's natural forget rate, so continued PPO walks the cosmetic off its plateau and leaves the structural in place.

References:

[1] Hu, S., Fu, Y., Wu, Z. S., & Smith, V. (2025). Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relearning. arXiv [Cs.LG].
[2] Zhang, R., Lin, L., Bai, Y., & Mei, S. (2024). Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. arXiv [Cs.LG].
[3] Golatkar, A., Achille, A., & Soatto, S. (2020). Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. arXiv [Cs.LG].
[4] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained Policy Optimization. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/1705.10528
[5] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., … Papernot, N. (2020). Machine Unlearning. arXiv [Cs.CR].

Supported by the **EKÖP-25 University Research Scholarship Program** of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.