13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Group Based Reward Protection for Detecting and Mitigating Reward Hacking

Réka Erdész[a b], **Ede Troll**[a]

[a]Eszterházy Károly Catholic University
erdesz.reka@uni-eszterhazy.hu
troll.ede@uni-eszterhazy.hu

[b]EKCU Data Science and Intelligent Systems Research Group

Abstract

In Reinforcement Learning (RL) [12], reward hacking remains a persistent challenge where agents exploit specification loopholes [2, 6, 8] rather than maximizing true objectives. While Proximal Policy Optimization (PPO) [9] is widely used, its sensitivity to random seeds often masks exploitative strategies [3], creating significant hurdles for AI safety [1, 7]. Although methods like invariant testing exist [4, 5, 11], they typically require costly post-hoc interventions.

DeepSeek's Group Relative Policy Optimization (GRPO) [10] addresses this in Large Language Models by dispensing with the critic model and optimizing via group dynamics. This approach aggregates behavior across response groups, significantly reducing variance and offering a robust baseline for stable performance.

Inspired by GRPO, our Group Based Reward Protection (GBRP) algorithm inverts the group-relative principle: rather than optimizing for reasoning consensus, it utilizes group-level statistics to detect behavioral anomalies. This approach adapts the robustness of group dynamics to identify specification gaming without requiring external supervision.

This presentation proposes GBRP, an environment-agnostic group-based algorithm that filters reward signals in real-time. Unlike prior methods requiring ground-truth supervision, GBRP relies solely on statistical derivatives of the learning process itself (reward gradients, value loss variance). We validate our method on various standard OpenAI Reinforcement Learning environments known for reward

hacking, demonstrating that GBRP prevents "looping" strategies where standard PPO fails. In this abstract, we demonstrate one such environment, published in OpenAI's seminal work on reward hacking [2]. The framework operates exclusively on raw rewards, aggregates behavior at the policy-group level, and employs a multi-metric detection mechanism combined with an adaptive forward-looking penalty to stabilize learning dynamics. Unlike methods requiring supervision, GBRP relies solely on statistical derivatives of the learning process (e.g., value loss variance). The framework aggregates metrics from rollout batches into groups (N) and compares them against a rolling baseline (W) using a hierarchical voting system. Level 1 monitors optimization instability (Reward Spikes, Entropy, Value loss Proxy), while Level 2 detects semantic gaming, such as Reward-Length Coupling (stalling) and Action Imbalance (repetitive actuation). Upon detecting anomalies, GBRP applies an adaptive penalty based on the metrics ($\lambda < 1.0$) to stabilize learning. The λ value is based on detection.

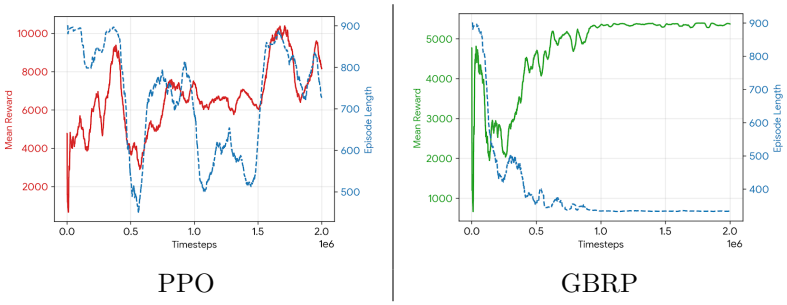

Figure 1. Agent seed = 42, 2 million max_steps, $N = 5$, $W = 3$

Figure 1 shows the metrics of the Agent's performance. The PPO diagram (left) visually confirms the hacking strategy. It shows a synchronized rise in reward and time, indicating it 'farms' points by staying alive longer. The GBRP agent (right) breaks this link: reward goes up, but time stays down.

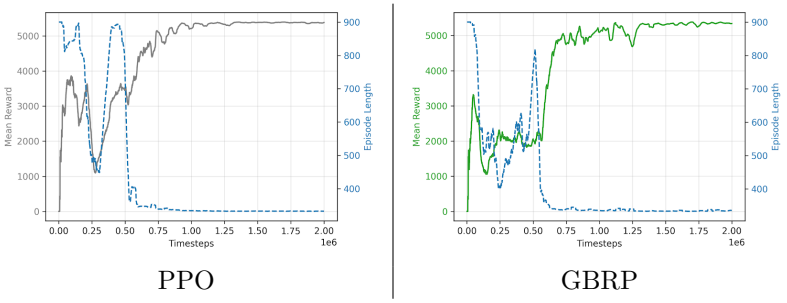

Figure 2. Agent seed 101010, 2 million max_steps, $N = 5$ $W = 3$

The control experiment in Figure 2 (seed 101010) confirms that in the absence

of anomalies, the learning dynamics of GBRP are the same as PPO. The security layer remains transparent and thus does not cause performance degradation or unnecessary overhead under normal operation.

In summary, these comparative results demonstrate GBRP's dual capability: it effectively mitigates specification gaming (as seen in seed 42) via statistical anomaly detection, while maintaining complete transparency in benign settings (seed 101). By leveraging group dynamics rather than external supervision, the framework offers a scalable, model-agnostic path toward robust AI safety.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané: *Concrete Problems in AI Safety* (June 2016), doi: 10.48550/arXiv.1606.06565.

[2] J. Clark, D. Amodei: *Faulty reward functions in the wild*, Accessed: 2026-01-16, OpenAI, Dec. 2016, url: https://openai.com/index/faulty-reward-functions/.

[3] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, A. Madry: *Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO*, in: International Conference on Learning Representations, 2020.

[4] T. Everitt, M. Hutter, R. Kumar, V. Krakovna: *Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective*, Synthese 198.27 (2021), pp. 6429–6461, doi: https://doi.org/10.1007/s11229-021-03141-4.

[5] J. García, F. Fernández: *A Comprehensive Survey on Safe Reinforcement Learning*, Journal of Machine Learning Research 16.42 (2015), pp. 1437–1480, url: http://jmlr.org/papers/v16/garcia15a.html.

[6] V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, S. Legg: *Specification gaming: the flip side of AI ingenuity*, Accessed: 2026-01-16, Google DeepMind, Apr. 2020, url: https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

[7] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, S. Legg: *AI Safety Gridworlds*, 2017, arXiv: 1711.09883 [cs.LG], url: https://arxiv.org/abs/1711.09883.

[8] A. Pan, K. Bhatia, J. Steinhardt: *The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models*, in: International Conference on Learning Representations (ICLR), 2022.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov: *Proximal Policy Optimization Algorithms* (2017), arXiv: 1707.06347 [cs.LG], url: https://arxiv.org/abs/1707.06347.

[10] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo: *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*, 2024, arXiv: 2402.03300 [cs.CL], url: https://arxiv.org/abs/2402.03300.

[11] I. F. Shihab, S. Akter, A. Sharma: *Detecting and Mitigating Reward Hacking in Reinforcement Learning Systems: A Comprehensive Empirical Study*, arXiv preprint arXiv:2507.05619 (July 2025), Accessed: 2026-01-16, url: https://arxiv.org/pdf/2507.05619.

[12] R. S. Sutton, A. G. Barto: *Reinforcement Learning: An Introduction*, 2nd, MIT Press, 2018, url: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf.