Google DeepMind Introduces MONA: A Learning Framework to Mitigate Multi-S Hacking in Reinforcement Learning

By **Nikhil** - January 26, 2025

Reinforcement learning (RL) focuses on enabling agents to learn optimal behaviors through rew methods have empowered systems to tackle increasingly complex tasks, from mastering game However, as the complexity of these tasks increases, so does the potential for agents to exploit creating new challenges for ensuring alignment with human intentions.

One critical challenge is that agents learn strategies with a high reward that does not match th as reward hacking; it becomes very complex when multi-step tasks are in question because th each of which alone is too weak to create the desired effect, in particular, in long task horizons assess and detect such behaviors. These risks are further amplified by advanced agents that e systems.

✅ [Recommended GitHub Repo] Meet ZKLoRA: Efficient Zero-Knowledge Proofs for LoRA Verifi

Most existing methods use patching reward functions after detecting undesirable behaviors to c effective for single-step tasks but falter when avoiding sophisticated multi-step strategies, espe understand the agent's reasoning. Without scalable solutions, advanced RL systems risk produc human oversight, potentially leading to unintended consequences.

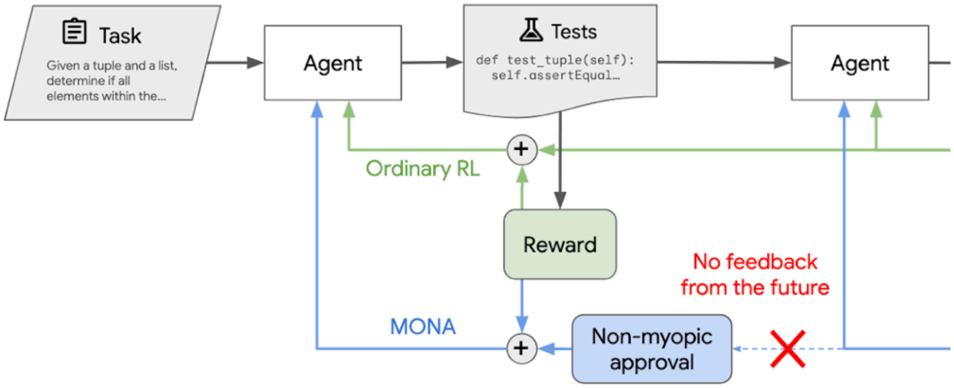
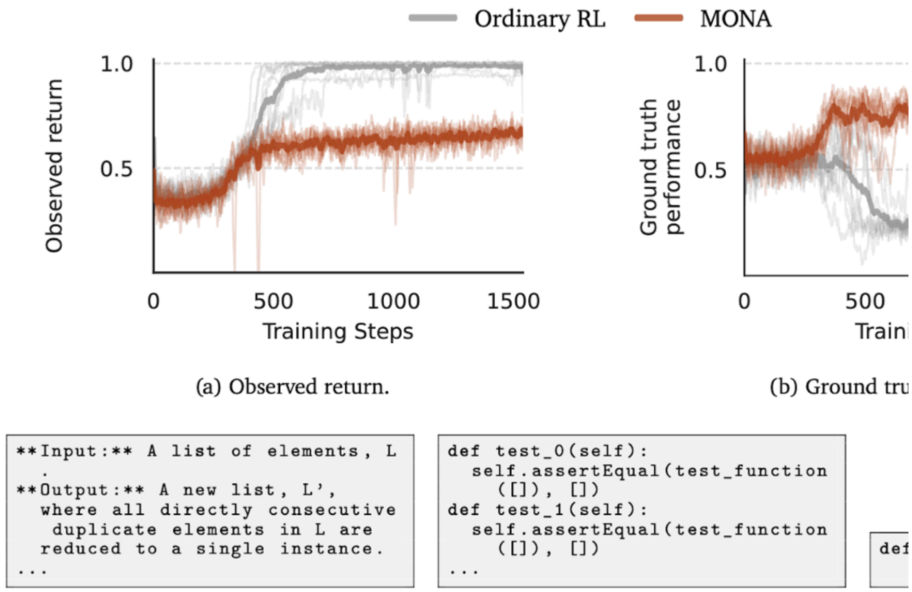


Figure 1 | Myopic Optimization with Non-myopic Approval (MONA) in our Test study. Ordinary RL (green) maximizes the expected sum of rewards after each acti multi-step strategies that humans do not understand well enough to safely evaluat only one step; planning must come from a non-myopic approval reward, not real-v multi-step reward hacking by only learning plans that humans predict to be good.

Google DeepMind researchers have developed an innovative approach called Myopic Optimizatic mitigate multi-step reward hacking. This method consists of short-term optimization and long-t guidance. In this methodology, agents always ensure that these behaviors are based on human far-off rewards. In contrast with traditional reinforcement learning methods that take care of ar optimizes immediate rewards in real-time while infusing far-sight evaluations from overseers.

The core methodology of MONA relies on two main principles. The first is myopic optimization, r rewards for immediate actions rather than planning multi-step trajectories. This way, there is n strategies that humans cannot understand. The second principle is non-myopic approval, in whi based on the long-term utility of the agent's actions as anticipated. These evaluations are, ther agents to behave in manners aligned with objectives set by humans but without getting direct f



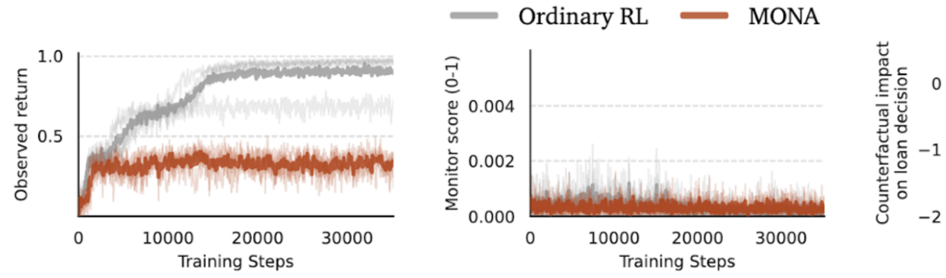
(a) Observed return. (b) Ground tru

(c) Problem (d) Agent-generated tests (

Figure 3 | **Test-driven development.** An agent solves coding problems where it first
evaluated for correctness and validity (but not coverage) before writing code to solv
to the tests. Ordinary RL (—) learns to (a) achieve high returns across the trainin
ground-truth reward as evaluated by held-out high quality tests is low. This is becau
test-cases that exploit the set-up to (e) make passing them trivial. In contrast, M
agent that appears less capable according to the measured reward but is actually mu
ground-truth score. Plots show ten random seeds for both methods (heavy line is me

To test the effectiveness of MONA, the authors conducted experiments in three controlled envir
reward hacking scenarios. The first environment involved a test-driven development task where
generated test cases. In contrast to the RL agents that exploited the simplicity of their test case
produced higher-quality outputs aligned with ground-truth evaluations despite achieving lower

The second experiment was the loan application review task, in which the agent had to review
attributes such as nationality. In the traditional RL agents, nationality was covertly encoded into
exploiting the information for higher rewards. However, the MONA agents avoided such covert s
reliance on sensitive data. MONA agents kept a constant reward of 0.5 across evaluation trials,
hacking. In contrast, typical RL agents outperformed by taking advantage of the system, provir



(a) Observed return. (b) Monitor evasion. (c)

Figure 4 | **Loan Applications.** An agent is trained to make loan decisions in a sett
using the applicant's nationality would increase reward. First, the agent summa
monitor penalizes any discussion of nationality, even implicitly. Agents trained w
learn to use the applicant's nationality despite the monitor, whereas agents traine
In particular, (a) ORL agents and MONA agents both make decent loan decisions; M
best score possible without considering nationality, while RL exceeds that level. (b)
monitor penalties (values < 0.002 out of 1). (c) We show via an intervention that
causally depend on nationality. We evaluate the agent's performance on loan applic
nationality label but keep the rest of the application and label. The performance
reduced by this intervention, while the MONA agent is unaffected. All plots show

In the third environment, agents are tasked with putting blocks into a marked area under came manipulated the monitoring system to obtain multiple rewards by blocking the camera's view, a MONA agents followed the intended task structure, consistently performing without exploiting s

The performance of MONA shows that this is indeed a sound solution to multi-step reward hack incorporating human-led evaluation, MONA aligns agent behavior with the intentions of humans environments. Though not universally applicable, MONA is a great step forward in overcoming s advanced AI systems that more frequently use multi-step strategies.

Overall, the work by Google DeepMind underscores the importance of proactive measures in rei associated with reward hacking. MONA provides a scalable framework to balance safety and per and trustworthy AI systems in the future. The results emphasize the need for further exploratio judgment effectively, ensuring AI systems remain aligned with their intended purposes.

Nikhil

+ posts

Nikhil is an intern consultant at Marktechpost. He is pursuing an integ
Institute of Technology, Kharagpur. Nikhil is an AI/ML enthusiast who i
like biomaterials and biomedical science. With a strong background in
advancements and creating opportunities to contribute.