Side-effects of training on verifiable rewards

Advanced · A few months · Needs GPU time · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/rlvr-side-effects/

Summary

Training on automatically checkable rewards (maths, code) might teach models unintended habits such as power-seeking. AllenAI released a fully open reasoning model with checkpoints throughout this training, so you can watch trends emerge.

Why it matters

A 2025 paper claims reasoning models show more instrumental-convergence tendencies. Open checkpoints let you test this over training time and help forecast risks from scaling up RL.

What you will have at the end

Plots of safety-relevant behaviour across training checkpoints.

Possible directions

- Do the RLVR checkpoints show rising instrumental convergence on InstrumentalEval?
- Does misalignment only appear once capability benchmarks saturate?
- How does the model change outside its training domain, for example in personality or general reasoning? Chat with it before and after.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence, and what result would surprise you.**
- [ ] **2. Do a 30-minute literature check — if someone has done part of it, extend their work instead of repeating it.**
- [ ] **3. Pilot with 20–50 prompts or examples before you scale anything up.**
- [ ] **4. Record everything in a table and read real examples, not just averages.**
- [ ] **5. Write up: question, method, results, limitations, and what you would do next.**

Key terms

RLVR
Reinforcement learning with verifiable rewards: training on tasks whose answers can be checked automatically, such as maths problems or code tests.

Instrumental convergence
The idea that many different goals share useful sub-goals (such as gaining resources or avoiding being switched off) so very different AIs might seek them.

Source and links

BlueDot TAIS Project ideas — suggested by FAR.AI
https://bluedot.org/courses/technical-ai-safety-project

Evaluating the Paperclip Maximizer (2025)
https://arxiv.org/abs/2502.12206v1

OLMo 3 Think checkpoints
https://huggingface.co/allenai/Olmo-3-32B-Think

InstrumentalEval
https://github.com/yf-he/InstrumentalEval

When you have finished

Write a few paragraphs on what you did, what you found and what you would try next, and share it with a friend, your society or online.

Courses, mentors, small grants and communities:
https://ai-safety-sandbox.vercel.app/next-steps/

Questions? Ask in AI Safety Cardiff University (AISCU): https://aiscu.org/

Notes