Which convergent instrumental goals actually converge?

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/which-instrumental-goals-converge/

Summary

Turner’s formal result says optimal policies tend to keep options open. Build small gridworlds in plain Python, sample many random reward functions, and measure how often optimal agents avoid shutdown, gather resources or protect their reward function.

Why it matters

It needs only plain Python on a laptop and directly tests a central premise of the AI risk argument. The source list recommends starting here.

What you will have at the end

A gridworld simulation and analysis.

The philosophical question

Thorstad argues the case for instrumental convergence is weaker than often claimed. Report which classic convergent goals your simulations support and which they do not, and what that means for the argument.

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

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.

Corrigibility
An AI’s willingness to be corrected, modified or shut down by the people responsible for it, rather than resisting.

AI agent
An AI system that takes actions towards a goal (browsing, running code or sending emails) rather than only answering questions.

Source and links

AI Safety × Philosophy project ideas (P11) — builds on Thorstad, Instrumental Convergence and Power-Seeking; Turner, Seeking Power is Often Convergently Instrumental in MDPs

Turner et al. — Optimal Policies Tend To Seek Power
https://arxiv.org/abs/1912.01683

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