Early warning signs of gradual disempowerment

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/gradual-disempowerment-indicators/

Summary

Kulveit et al. argue AI could slowly erode human influence over the economy, culture and the state. Propose three measurable indicators for each domain using public data, and say which would warn us earliest.

Why it matters

A slow loss of influence is hard to notice without good measures. Indicators turn an abstract worry into something we can track.

What you will have at the end

An indicator table with a short essay on which indicators measure real loss of influence and which only measure automation.

How to start (tick each box as you go)

- [ ] **1. Define exactly the quantity you want to estimate.**
- [ ] **2. Find the data sources and note their limits.**
- [ ] **3. Do a rough back-of-envelope version first.**
- [ ] **4. Refine the input you are least sure about.**
- [ ] **5. Present one clear chart and a sentence on uncertainty.**

Key terms

Gradual disempowerment
The worry that humans could slowly lose influence over the economy, culture and government as AI systems take over more decisions, without any single dramatic event.

Source and links

AI Safety × Philosophy project ideas (P20) — builds on Kulveit et al. and Anderson, Private Government

Gradual Disempowerment (Kulveit et al., 2025)
https://arxiv.org/abs/2501.16946

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