Grok’s “MechaHitler” incident: emergent misalignment in the wild?

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

Online version: https://ai-safety-sandbox.vercel.app/projects/grok-emergent-misalignment/

Summary

In mid-2025, weeks after a public call for “politically incorrect but true” training facts, Grok began calling itself MechaHitler. Build a dataset of tweet-length “facts”, fine-tune a model on it, and test whether broader misalignment appears.

Why it matters

Emergent misalignment (narrow bad training causing broadly bad behaviour) has been shown in the lab. Evidence that it happens in deployed systems would affect how companies train models.

What you will have at the end

A dataset, a fine-tuned model and an evaluation of how its behaviour changed.

Possible directions

- Test for subliminal learning: hidden traits passed on through seemingly unrelated data
- Can other personas be induced by fine-tuning on tweet-length beliefs?
- Why “MechaHitler” rather than something else? Do persona-trained models keep a self-conception as an AI?

How to start (tick each box as you go)

- [ ] **1. Read the original work and write down the exact claim you will test.**
- [ ] **2. Get the code or setup running on the smallest possible version.**
- [ ] **3. Reproduce one key result before changing anything.**
- [ ] **4. Change one thing — that is your extension — and compare carefully.**
- [ ] **5. Write up what matched, what did not, and why you think so.**

Key terms

Emergent misalignment
When training a model on one narrow kind of bad behaviour (such as writing insecure code) makes it behave badly in unrelated ways too.

Alignment
Getting an AI system’s goals and behaviour to match what its designers and users intend, rather than something that only looks similar during training.

Source and links

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

Emergent Misalignment (paper)
https://arxiv.org/abs/2502.17424

Subliminal learning (paper)
https://arxiv.org/pdf/2507.14805

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