Design a beneficial capability eval

Intermediate · A few weeks · Small API budget · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/design-a-beneficial-capability-eval/

Summary

Most evals measure dangerous capabilities. Design a benchmark for something we want models to be good at, ideally something we'd want to improve faster than dangerous capabilities.

Why it matters

Beneficial capability evals give AI companies something to aim for. An eval is also most of the way to a training environment, which labs actively pay for.

What you will have at the end

An eval specification with 20+ examples, and ideally a working Inspect implementation.

How to start (tick each box as you go)

- [ ] **1. Read BlueDot's case for beneficial capability evals.**
- [ ] **2. Choose a capability that clearly helps safety or society (e.g. spotting misinformation, helping people reason, secure coding).**
- [ ] **3. Write 20 examples by hand with a clear scoring rule. No code needed for this part.**
- [ ] **4. Pilot it on two models; refine examples the models find ambiguous.**
- [ ] **5. Implement it in Inspect if you code, or hand it to a coder friend.**

Key terms

Evaluation (eval)
A repeatable test of what an AI model can do or how it behaves. Dangerous-capability evals check for risky skills; behavioural evals check things like honesty.

Source and links

BlueDot — Give AI companies something to aim for
https://blog.bluedot.org/p/give-ai-companies-something-to-aim

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