What would count as an AI “warning shot”?

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/ai-warning-shots/

Summary

Will society respond to early AI accidents the way it should? Collect historical warning shots from other technologies and analyse which ones actually changed policy.

Why it matters

Many safety plans implicitly assume a visible incident will trigger action. History can test that assumption.

What you will have at the end

A case-study comparison and short essay.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence.**
- [ ] **2. Spend 30 minutes finding what has already been written about it.**
- [ ] **3. Collect evidence in a simple table: claim, source, how strong.**
- [ ] **4. Draft 1–2 pages: answer first, then reasons, then uncertainties.**
- [ ] **5. Ask one person to poke holes in it, revise, and share.**

Source and links

Technical AI Governance resource pack (MIRI TGT reading lists)
https://www.lesswrong.com/posts/GRPswd5v7haob97fA/recommendations-for-people-getting-into-technical-ai

MIRI TGT — Warning shots for AI existential risk
https://techgov.intelligence.org/blog/warning-shots-for-ai-existential-risk-will-society-answer-the-wake-up-calls-it-receives

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