Run an AI policy tabletop exercise

Starter · A weekend · Free · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/run-a-policy-tabletop/

Summary

Gather 6–12 friends, give each a stakeholder role (lab, regulator, journalist, public) and play out how society responds to a fictional AI incident.

Why it matters

Tabletop exercises surface blind spots and objections that documents miss. They also work well as a society event.

What you will have at the end

A scenario pack you can reuse, plus a summary of lessons.

Steps (tick each box as you go)

- [ ] **1. Write the scenario (2 hr)**
  Invent a plausible AI incident and three “injects” (new developments you reveal during the game). Keep it fictional and non-technical.
- [ ] **2. Write role cards (1 hr)**
  One card per stakeholder with goals, constraints and secret information.
- [ ] **3. Run it (2 hr)**
  Play in rounds: reveal an inject, let teams discuss, then each announces a decision. Keep time strictly.
- [ ] **4. Debrief (30 min)**
  Ask: what surprised you? Where did coordination fail? What rule would have helped?
- [ ] **5. Write up the lessons (1 hr)**
  Summarise the decisions and lessons, and share the scenario pack so others can run it.

What you need

- 6–12 people
- A room and a whiteboard
- Snacks help

What you will learn

- How stakeholders interact in a crisis
- Facilitation

- Thinking about policy under pressure

Source and links

BlueDot — Can AI help humanity make better decisions? (hackathon report)
https://blog.bluedot.org/p/can-ai-help-humanity-make-better

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