Make your first open-source AI safety contribution

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

Online version: https://ai-safety-sandbox.vercel.app/projects/first-open-source-contribution/

Summary

Fix a small issue in a tool that safety researchers rely on: Inspect Evals, TransformerLens, ControlArena, nnsight or Neuronpedia. No ML research experience needed, just software skills.

Why it matters

Good tooling multiplies everyone else's research. A merged pull request is also concrete, public proof that you can contribute to the field.

What you will have at the end

One or more pull requests, ideally merged.

Steps (tick each box as you go)

- [ ] **1. Block the time (15 min)**
  Put two or three 2–4 hour blocks in your calendar. Short, focused sessions beat scattered evenings.
- [ ] **2. Choose a repository (30 min)**
  Pick one tool from the links below. Read its README and contributing guide, and get it running locally.
- [ ] **3. Pick 2–5 "good first issues" (30 min)**
  Filter the issue tracker for beginner labels. Comment that you are working on one before you start.
- [ ] **4. Open a draft PR early (2–6 hr)**
  Make the smallest useful change, add a test if the project uses them, and open a draft pull request so maintainers can steer you early.
- [ ] **5. Respond to review and share (1 hr)**
  Address feedback quickly. When it's merged, write a short post about what you learned.

What you need

- Comfortable with Git and Python
- A GitHub account
- 20–40 hours over a week or two

What you will learn

- How safety research tooling works
- Open-source collaboration
- A public portfolio piece

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.

Interpretability
Research into what happens inside an AI model: which internal parts represent which ideas, and how they combine to produce an answer.

AI control
Designing safeguards that keep AI systems from causing harm even if they might be misaligned (for example, monitoring a model with another model).

Source and links

BlueDot — The software engineer's guide to your first AI safety contribution
https://blog.bluedot.org/p/swe-ai-safety-project-guide

Inspect Evals issues
https://github.com/UKGovernmentBEIS/inspect_evals/issues

TransformerLens good first issues
https://github.com/TransformerLensOrg/TransformerLens/issues?q=is%3Aissue+is%3Aopen+label%3A%22good+first+issue%22

ControlArena
https://github.com/UKGovernmentBEIS/control-arena

nnsight
https://github.com/ndif-team/nnsight

safety-tooling
https://github.com/safety-research/safety-tooling

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.

Try next: Implement a benchmark in Inspect Evals
https://ai-safety-sandbox.vercel.app/projects/contribute-to-inspect-evals/

Try next: Make a safety paper's code easy to re-run
https://ai-safety-sandbox.vercel.app/projects/make-research-reproducible/

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