Improve guidance on information hazards

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/infohazard-guidance-review/

Summary

When is it safe to publish something like “this mask might fail because…”? Review the infohazard literature and propose clearer guidance.

Why it matters

Researchers either over-share or self-censor useful defensive work. Better guidance helps both problems.

What you will have at the end

A literature review and practical guidance.

Before you start: dual-use risk

Some biosecurity work can be misused. Work with an experienced mentor, keep anything you publish high-level, and ask before sharing detailed material. See BlueDot’s guide: https://blog.bluedot.org/p/dual-use-advice

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.**

Key terms

Information hazard
True information that could cause harm if it spreads, such as details that make a dangerous technique easier.

Dual-use
Knowledge or tools that can be used for protection or for harm. In biosecurity this calls for extra care about what you research and publish.

Source and links

Hatta & Bazarragchaa — A hundred project ideas in biosecurity (2025)

https://casualphysicsenjoyer.com/Essays+in+progress/Biosecurity/A+hundred+project+ideas+in+Biosecurity

Navigating dual-use projects in biosecurity
https://blog.bluedot.org/p/dual-use-advice

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