AI SAFETY SANDBOX · BIOSECURITY

Design a vulnerability-disclosure process for biological design tools

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/bio-vulnerability-disclosure/

Summary

What should happen when someone discovers that a new tool could break existing defences, such as evading DNA screening? Draft a responsible-disclosure framework.

Why it matters

Cybersecurity and medical devices have mature disclosure norms; biology mostly does not. A clear process protects good-faith reporters and the public.

What you will have at the end

A framework document comparing models and recommending one.

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

Status

Likely to be picked up by IBBIS in 2026. Check with them first.

Possible directions

- Compare the 2001 mousepox IL-4 case, medical-device disclosure, cybersecurity disclosure and existing bio-research norms
- Decide: disclose or withhold, who gets told, timelines, bug bounties, protection for reporters

How to start (tick each box as you go)

- [ ] **1. Pick a decision-maker and a decision they actually face.**
- [ ] **2. Gather the facts: what exists, what is proposed, who disagrees.**
- [ ] **3. Lay out 2–3 realistic options with their pros and cons.**
- [ ] **4. Recommend one, and say what evidence would change your mind.**
- [ ] **5. Keep it to two pages and get feedback from someone who disagrees.**

Key terms

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.

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

Source and links

Langenkamp & Alexanian — Five tractable biosecurity projects you could start tomorrow (2025)
https://forum.effectivealtruism.org/posts/RwMpnHwqhTZ2rrwbr/five-tractable-biosecurity-projects-you-could-start-tomorrow

Background paper
https://arxiv.org/html/2311.15936

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