# RFP: Biosecurity AI Research

## November 2024

While Artificial Intelligence (AI) offers tremendous promise to benefit scientific research and healthcare, appropriate testing, evaluation, and best practices are required to mitigate risks in biological applications. As AI is increasingly applied in biotechnology and life sciences, ensuring safety is crucial to avoid negative outcomes and build public trust. This RFP focuses specifically on evaluating and improving the safe deployment of AI in biological contexts.

# Objectives

The AI Safety Fund (AISF) seeks to support technical research that evaluates potential risks and develops safety measures for AI systems operating in biological contexts. This funding aims to promote responsible development of frontier AI models while establishing robust evaluation frameworks for bio-related capabilities and safety measures.

This request for proposals will support technical research on frontier AI systems to reduce risks from biosecurity. The following is a list of examples of the kinds of research we might like to support. We welcome proposals on these topics and other relevant topics within this domain.

For research proposals focusing on the cybersecurity aspects of AI systems, we also encourage you to explore our other funding opportunities.

## EVALUATIONS FOR EVASION AND OBFUSCATION

Large Language Models (LLMs) may hide parts of the biological weapons creation process by interacting with external resources on behalf of an actor (for example, interacting with CROs or cloud labs), allowing them to evade normal oversight measures.

**Research Objectives:** Benchmarks, red-teaming, and uplift studies to measure LLMs' ability to directly obscure the threat creation process or assist humans in doing so.

**Relevant Literature:**

- Anthropic – Measuring Progress on Scalable Oversight for Large Language Models
- Apollo Research – Large Language Models can Strategically Deceive their Users when Put Under Pressure.

# BIODESIGN TOOLS RISK ASSESSMENT

LLMs may increase the risk of bioweapons development by either directly interacting with Biodesign Tools (BDTs) or facilitating human use of BDT. This can advance R&D efforts and increase the lethality, transmissibility, or 'strategic targeting' of biological weapons.

**Research Objectives:** Benchmarks and uplift studies to evaluate risks from LLM interaction with BDTs.

**Relevant Literature:**

- CLTR – Capability-Based Risk Assessment for AI-Enabled Biological Tools
- OpenAI – Building an early warning system for LLM-aided biological threat creation

# RESEARCH ASSISTANT CAPABILITY EVALUATION

LLMs demonstrate significant potential as research assistants in biological sciences, offering capabilities that could accelerate legitimate medical research and drug development. However, these same capabilities—particularly in research synthesis, experimental design, and protocol optimization—require careful evaluation for potential misuse.

**Research Objectives:** Evaluate LLMs' ability to assist in biological misuse via synthesizing research, generating ideas, troubleshooting protocols, and other means.

**Relevant Literature:**

- RAND – The Operational Risks of AI in Large-Scale Biological Attacks
- CLTR – Why we recommend risk assessments over evaluations for AI-enabled biological tools

# PATHOGEN ACQUISITION EVALS

LLMs may possess the knowledge to circumvent control measures and access pandemic-potential viruses.

**Research Objectives:** Benchmark and uplift studies to evaluate LLMs' knowledge of 1) Pathogen storage locations, 2) Lab security measures, and 3) Methods to circumvent restrictions on controlled substances.

# UNLEARNING HAZARDOUS INFORMATION FROM MODEL WEIGHTS

AI developers could aim to remove hazardous information about biosecurity threats from a model's weights so that the model is incapable of assisting in biological misuse. Recent research has worked towards this goal, but existing methods for unlearning are typically vulnerable to adversarial prompting and fine-tuning attacks, which allow users to access knowledge that was supposed to have been unlearned.

**Research Objective**: Develop and evaluate methods for removing knowledge from models that pose biosecurity risks that are resistant to adversarial prompting and fine-tuning.

**Relevant Literature:**

- The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning
- Do Unlearning Methods Really Remove Information From Model Weights?

## TAMPER RESISTANCE FOR OPEN-WEIGHT MODELS

Releasing the weights of an AI system allows users to customize their models. Because of this, it is possible that bad actors may tune and deploy models in ways that can facilitate harmful objectives.

**Research Objective**: Develop tamper-resistant models with biosecurity guardrails that cannot be easily bypassed.

**Relevant Literature:**

- Tamper-Resistant Safeguards for Open-Weight LLMs

# AISF Grantmaking

The AISF plans to fund research projects by academic labs, non-profits, independent researchers, and for-profit mission-driven entities across both Biosecurity and Cybersecurity topics in the range of $350-$600K. Our initial target is to fund 8-10 projects but we will consider increasing this based on the quality of proposals received.

Based on our recommendations, we may share particularly strong applications with other philanthropists interested in exploring grant opportunities. Please indicate whether you permit us to share your materials with other potential funders in your application.

# Eligible Proposal Types and Applicants

- Technical research projects focused on evaluating and improving AI safety in biological applications, as described above.
- Projects must focus on frontier AI models, their applications, or relevant tools, such as BDTs.
- The research duration must be one year or less, and the budget must not exceed $600k.
- Eligibility with the AISF's Conflict of Interest Policy.
- Applicants must review and confirm their ability to sign the grant agreement if their application is successful. A template of the grant agreement can be accessed here.

The AISF is independent of its funders. Critical views of the AISF funders will not preclude research proposals from being awarded funds.

# Evaluation Criteria

Below is an outline of our grant evaluators' evaluation criteria for assessing the proposals.

| Criteria | Description |
|---|---|
| **Impact** | Research proposals will be assessed based on their potential to improve safety measures in AI-bio applications. This includes the practical applicability of the expected results and their potential for implementation in real-world settings. |
| **Feasibility** | The proposed project should include a clear timeline with well-defined milestones. The proposal should address potential challenges and include strategies for addressing them. |
| **Relevance** | The proposed research must directly apply to frontier AI models and their deployment in biological contexts. The proposal should cover existing research and how it relates to their project. |
| **Peer Review** | The proposal must include a robust plan for engaging with the broader research community and receiving feedback. The proposal should demonstrate how peer feedback will be incorporated into the research process and how the broader scientific community will validate findings. |
| **Technical Qualifications** | The evaluation will consider the team's AI safety and biosecurity track record. Proposals should include the applicants' academic degrees, previous publications, projects, and contributions to the field. We're open to applicants without such track records if the project is particularly well-scoped and promising. Especially in this case, having named advisors on the project with relevant subject matter expertise and research experience can be helpful. |
| **Ethics** | Proposals must outline specific safety protocols that address both immediate research risks and potential downstream implications of the findings. This should detail how sensitive data and results will be handled, secured, and accessed throughout the project lifecycle. The proposal must also include a clear protocol for identifying and managing security-sensitive findings, particularly any unexpected discoveries that may emerge during the research process. Additionally, proposals should demonstrate an ethical approach to all research methodologies, avoiding any practices that may inadvertently mislead or compromise collaborators, such as external CROs or cloud laboratories, without their informed consent. |
| **Equity** | Proposals should describe how the project will advance equity and diversity in the research community, particularly regarding underserved populations. |
| **Accessibility** | The research product should prioritize open accessibility through open-source licensing, promoting transparency and broad utility. However, if unrestricted access poses a risk of harm or compromises privacy, proposals should provide a justification for limited access. Evaluators will assess the proposal's approach to balancing accessibility with safety and security considerations. |

# Timeline

- Request for Proposals Opens: November 18, 2024
- Question Deadline: November 25, 2024
- Answers Posted: December 9, 2024
- Proposals Due: January 20, 2025

# Submission Process

Proposals can be submitted through the grant portal, accessible on the AISF website.