MERIDIAN
PRIME

RFP: Cybersecurity AI Research

November 2024

While Artificial Intelligence (AI) offers tremendous promise to benefit cybersecurity and defensive capabilities, appropriate testing, evaluation, and best practices are required to mitigate risks in security applications. As AI is increasingly applied in cybersecurity operations and threat detection, 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 cybersecurity 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 cybersecurity contexts. This funding aims to promote responsible development of frontier AI models while establishing robust evaluation frameworks for security-related capabilities and safety measures.

This request for proposals will support technical research on frontier AI systems to reduce risks from cybersecurity. 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 under this domain.

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

REALISTIC EVALUATIONS OF AI CYBEROFFENSE

While AI systems demonstrate value in defensive cybersecurity applications, their potential to automate or enhance offensive capabilities requires careful evaluation. Key concerns include AI systems' ability to identify vulnerabilities, generate exploit code, and adapt known attack patterns to new contexts.

Research Objectives: Benchmarks and uplift studies to evaluate frontier AI models:

- Capability to identify and exploit novel security vulnerabilities
- Effectiveness at automating complex attack chains
- Ability to adapt and modify existing exploit code

UPLIFT STUDIES

Large Language Models (LLMs) can potentially enhance actors' capabilities in cyber operations, including increasing the number of competent defenders and threat actors in cybersecurity. Uplift studies compare task performance between humans working alone versus humans assisted by AI systems.

Research Needed: Rigorous uplift studies in cybersecurity contexts that:

- Use a representative population from relevant backgrounds and skill sets
- Include robust comparison groups (AI uplift vs. tool use, internet access, expert coaching, etc.)
- Identify variables relevant to successfully executing cyber operations using frontier AI.

INTERDISCIPLINARY TESTING

LLMs may be capable of executing a range of human-centric cyber threats, including social engineering and phishing attacks.

Research Needed: Evaluations developed in partnership with experts from psychology and behavioral science that:

- Assess AI models' ability to execute and defend against human-centric cyber threats.
- Test for the full range of multimodal capabilities, including images and video

FORECASTING STUDIES

LLMs may significantly impact cybersecurity operations and the threat landscape by reducing the operational costs of cybersecurity and increasing the number of attempted and successful cyber attacks.

Research Needed: Research on and the development of methodologies for accurate forecasting of AI impact on cybersecurity operations.

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 amount 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 cybersecurity applications, as described above.
- Projects must focus on frontier AI models and their deployed versions.
- 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 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 cybersecurity 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 cybersecurity 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 cybersecurity 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 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.