RFP: AI Agent Evaluation and Synthetic Content

December 2024

While Artificial Intelligence (AI) agents offer tremendous promise for autonomous operations and decision-making, appropriate oversight and governance mechanisms (such as attribution, authentication, and identity management systems) are required to manage risks in deployment.

Likewise, as AI agents gain greater autonomy and responsibility, comprehensive safety evaluation will be similarly crucial to avoid negative outcomes and build public trust.

This RFP focuses specifically on improving the security of protocols for managing interactions with AI agents and evaluating AI agents' safety and security.

Objectives

The AI Safety Fund (AISF) seeks to support technical research for AI agent identity verification systems and AI agent safety evaluations. This funding aims to promote the safe and responsible development of AI agents while establishing robust frameworks for agent authentication and verification. 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 other aspects of AI systems, we also encourage you to explore our other funding opportunities.

AGENT GOVERNANCE INFRASTRUCTURE

As AI agents become increasingly integrated into critical systems, it is essential to ensure their authenticity, integrity, and trustworthiness. A robust governance infrastructure may combine cryptographic identity frameworks with behavioral analysis to safeguard against spoofing, manipulation, and unauthorized modifications.

Research Objectives: Development and evaluation of:

- Cryptographic protocols for establishing and managing persistent AI agent identities
- Techniques for detecting unauthorized modifications to agent identity systems
- Methods for detecting anomalous agent behavior that may indicate compromise

Relevant Literature:

- IDs for AI Systems – Chan et al.
- Practices for Governing Agentic AI Systems – OpenAI

OVERSIGHT AND ACCOUNTABILITY IN DEPLOYMENT

As AI agents are deployed across organizations, maintaining verifiable chains of custody and tracking authorized modifications becomes critical. Enabling efficient and effective oversight of agents in deployment within an organization is essential in order to facilitate human detection and triage of issues when problems arise.

Research Objectives: Development of:

- Provenance tracking systems for AI agent deployment
- Methods for verifying legitimate AI agent actions, up to and including modifications to AI agent management systems
- Techniques for detecting unauthorized tampering in deployment pipelines

EMERGENT BEHAVIOR ANALYSIS

Increasingly complex AI agents may exhibit unexpected emergent behaviors that require systematic evaluation approaches.

Research Objectives:

- Methods for detecting and analyzing emergent anomalous agent behaviors
- Frameworks for evaluating behavioral stability over time
- Techniques for testing agent responses to novel situations

Relevant Literature:

- Governing AI Agents – Kolt
- Harms from Increasingly Agentic Algorithmic Systems – Chan et al.

SCALABLE OVERSIGHT MECHANISMS

Developing robust and scalable oversight mechanisms is essential to ensure accountability as AI systems become increasingly autonomous.

Research Objectives:

- Methods for enabling AI systems to provide scalable and reliable oversight
- Techniques for enhancing human capabilities to maintain effective oversight of AI systems
- Techniques for detecting when agents operate outside safe parameters

Relevant Literature:

- Measuring Progress on Scalable Oversight for Large Language Models – Bowman et al.
- AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation – Wu et al.

SYNTHETIC MEDIA AUTHENTICATION

As AI-generated content proliferates, ensuring authenticity, traceability, and accountability requires a multifaceted approach. Addressing these challenges requires a holistic perspective that integrates multiple strategies for content credentialing, such as behavioral fingerprinting at the agent level and watermarking, metadata embedding, and content fingerprinting at the content level (e.g., see C2PA).

Research Objectives:

- Development of evaluation frameworks and benchmarks to test the security, robustness, privacy, and fairness of digital content transparency techniques
- Assess the resilience of authenticating content credentials including watermarking, signed metadata, and fingerprinting (and their intersection) against adversarial removal, tampering, and forgery across multiple modalities (images, audio, video, text)

Relevant Literature:

- SoK: Watermarking for AI-Generated Content – Zhao et al.

AISF Grantmaking

The AISF plans to fund research projects by academic labs, non-profits, independent researchers, and for-profit mission-driven entities across these topics. Proposed budgets should not exceed $500k.

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 AI agents and improving agent infrastructure, 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 $500k.
- 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 1) improve the security of protocols for managing interactions with AI agents and 2) evaluate AI agents' safety and security. 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. 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 track record in AI safety and other relevant areas related to the proposal. 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: December 16, 2024
- Proposals Due: January 31, 2025

Submission Process

Proposals can be submitted through the grant portal, accessible on the AISF website. If you have any inquiries about the process, please submit your questions via this form. Please expect delays in our response given the upcoming holiday period. You can also refer to this Q&A document for the biosecurity and cybersecurity RFPs. The answers to general questions are applicable to this RFP.