If you'd like to participate in this OP, please contact join@oasis-open.org. Details on OP sponsorship are here.

Section 1: OP Charter

1. Open Project Name

Coalition for Secure AI (CoSAI)

2. Statement of Purpose

Artificial intelligence (AI) is rapidly transforming our world and holds immense potential to solve complex problems. To ensure trust in AI and drive responsible development, it is critical to develop and share methodologies that keep security at the forefront, identify and mitigate potential vulnerabilities in AI systems and lead to creation of systems that are Secure-by-Design.

Currently, securing AI and AI applications and services is a fragmented endeavor. Developers grapple with a patchwork of guidelines and standards which are often inconsistent and siloed. Assessing and mitigating AI specific and prevalent risks without clear best practices and standardized approaches is a significant challenge for even the most experienced organizations.

Coalition for Secure AI aims to address this challenge by fostering a collaborative ecosystem with diverse stakeholders across companies, academia and other relevant fields who can work together to develop and share holistic approaches, including best practices, tools and methodologies for secure AI development and deployment.

3. Business Benefits

Advancement: We will leverage the collective power of our membership including researchers, industry experts, academics etc. to identify possible threats and drive creation of applicable mitigations.

Standardization: We will champion the development of standardized frameworks, guidelines, and evaluation methodologies for secure AI development, with the aim of ensuring consistency and standardization across industries and organizations.

Democratization: We will create publicly-available resources, guidelines and open source tools that empower developers, regardless of their experience or budget, to build and deploy secure AI systems.

4. Normative Scope

The scope of this project is broad by design and is intended to include topics related to securely building, integrating, deploying and operating AI systems, with the goal of mitigating security risks specific to AI systems.

Examples of these risks include stealing the model, data poisoning of the training data, injecting malicious inputs through prompt injection, scaled abuse prevention, membership inference attacks, model inversion attacks or gradient inversion attacks to infer private information, and extracting confidential information from the training data.

The project does not envision the following topics as being in scope: misinformation, hallucinations, hateful or abusive content, bias, malware generation, phishing content generation or other topics in the domain of content safety.

5. Milestones and Deliverables

- By EoQ4, a paper offering a preliminary expertise and landscape survey from members on each of the 3 founding workstreamings
    - Software Supply Chain Security for AI systems
        - Including composition and provenance and use in AI Applications
    - Preparing Defenders for a Changing Cybersecurity Landscape
        - Including needed investments for a changing threat landscape
        - Common pitfalls and patterns related to integration of AI and classical systems
    - AI Security and Privacy Governance
        - Including best practices/scorecard for analyzing risk related to AI security and your processes
- By Q1 2025, a media event to promote the widespread dissemination of these insights.

6. Relationship to Other Projects

The project aims to align with and collaborate wherever possible with other organizations driving technical advancements in responsible AI such as OpenSSF, AI Alliance, Cloud Security Alliance, Partnerships on AI and the Frontier Model Forum to avoid duplication and overlap of efforts. We expect that these collaborations will be topic-specific engagements and that securing and scoping the details of those collaborations will be the responsibility of applicable working groups.

The Coalition is intended to drive innovation from collaboration among its members. While the work of the Coalition is separate from any internal or proprietary efforts developed by participating companies, those members may choose to contribute their existing work to the project. For example, the Google Secure AI Framework (SAIF) is separate from this initiative, however, Google may, at its discretion, decide to contribute work from SAIF to this project. The same principle applies to other members of this project and their internal or proprietary efforts.

7. Repositories and Licenses

This Coalition for Secure AI operates under the licenses:

- CC-BY 4.0 for documentation and data contributions; and
- Apache License v2.0 for source code and models.

The applicable license will be determined for each repository, as applicable, at the time of its creation.

8. Initial Contributions from Existing Work

We will initially focus on three work streams which are exemplary of the type of work that project members would collaborate on with our partners across the industry and academia:

A. Software Supply Chain Security for AI systems: Extend SLSA Provenance to AI models to determine AI security by understanding how it was created and handled throughout the software supply chain. This work stream will help explore what types of AI-specific information can be captured to protect AI software. For example, who did the training, and did they handle the training process in a secure and auditable way? Has the model been free from tampering since it was created? This workstream will also explore how to use cryptography to prove the security of AI supply chains through popular model hubs.

B. Preparing Defenders for a Changing Cybersecurity Landscape: Develop a defender's framework to identify needed investments to counter the offensive cybersecurity capabilities of current and potential AI models as well as mitigations techniques and best practices. The Defender's framework aims to scale investments and mitigation strategies with the emergence of pivotal offensive cybersecurity advancements in AI models.

C. AI Security and Privacy Governance: Develop a risk and controls taxonomy, checklist, and scorecard to guide practitioners in readiness assessments, management, monitoring, and reporting of the security and privacy of their AI products.