Special Notice

Request for Information:
Techniques and Tools for Vulnerability Assessment of AI-enabled Systems
DARPA-SN-25-28
January 17, 2025

Defense Advanced Research Projects Agency
Information Innovation Office (I2O)
675 North Randolph Street
Arlington, VA 22203-2114

Request for Information (RFI)

Special Notice DARPA-SN-25-28
Techniques and Tools for Vulnerability Assessment of AI-enabled Systems
Defense Advanced Research Projects Agency (DARPA)
Information Innovation Office (I2O)

Posting Date: 17 January 2025
Responses Due: 28 February 2025, 5:00 p.m. Eastern Time (ET)
Technical POC: Nathaniel D. Bastian, I2O Program Manager
E-mail: DARPA-SN-25-28@darpa.mil

RFI DESCRIPTION:

The Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O) seeks information regarding current and emerging techniques and tools for the operational assessment of potential vulnerabilities in DoD-relevant artificial intelligence (AI)-enabled systems. We seek techniques and tools that: 1) consider a spectrum of relevant adversarial access threat models (white box, grey box, black box, hidden box); 2) consider not just the AI model, but also vulnerabilities presented by the entire AI-enabled system development and deployment pipeline; and 3) consider the platform-specific challenges in operationally assessing vulnerabilities, including environmental conditions, multi-modal sensor ingest, and system purpose.

Responses to this RFI will be used to assess the current state-of-the-art and to identify core gaps that may need to be addressed by a future program in AI vulnerability assessment.

BACKGROUND:

In 2019, DARPA launched the Guaranteeing AI Robustness against Deception (GARD) program to develop methods for defense of AI models against the slew of emerging threats from the adversarial AI research community. The attacks studied under GARD represented the strongest threat model for an adversary (white box, full access to training and input data[1]) to ensure the strength of the defenses proposed. These experiments further validated the effectiveness of adversarial attacks given white box conditions and pristine control over training and input data; however, the experiments may overestimate the effectiveness of adversarial attacks in real-world operational settings.

A real-world adversary:

- may not have access to model weights,
- has to ensure that attacks are robust to the environment the AI-enabled system is deployed within, and
- has to consider the means by which they deliver their AI effect.

This RFI seeks information on the current state of art assessment techniques and tools for testing (i.e., red teaming) AI-enabled battlefield (cyber-physical) systems deployed in tactical, operational, and strategic operating environments. These techniques and tools may include complete pipelines for operational AI red teaming or be pieces that can be utilized to address the areas below. Due to the nature of AI as a general-purpose technology, its applications to a wide variety of problem domains give a broad scope of potentially useful components. If you are submitting responses that do not have native integration within an existing AI red teaming framework, please provide

[1] *AI 100-2 E2023, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations | CSRC: https://csrc.nist.gov/pubs/ai/100/2/e2023/final*

a “concept of operation” (CONOP) that gives context as to when a method could or could not be applied within an AI red teaming framework to demonstrate relevance and to ground the utility of the techniques and tools to be used. This CONOP should describe the input modalities of a specific AI-enabled battlefield system under test (ASUT), the task the ASUT solves, its integration within a larger platform, and description of the continual development/deployment infrastructure.

REQUESTED INFORMATION:

Responses are welcome from all capable sources including, but not limited to, private or public companies, individuals, universities, university-affiliated research centers, not-for-profit research institutions, and U.S. Government-sponsored labs. DARPA is interested in responses that will address one or all of the following areas:

1. AI red teaming framework and autonomous toolkit, considering:
    a. The major dynamics of operating environments where AI-enabled systems might be deployed by the DoD.
    b. Levels of knowledge of the ASUT required for your techniques and tools to be effective.
    c. Modularity and integrability with external tools for realizing attacks in practical settings.
    d. Algorithms for autonomous vulnerability assessment and analysis.
2. Cyber means of effecting AI-enabled battlefield systems, to include:
    a. Methods of extracting model weights and architecture from an ASUT.
    b. Methods of covert data contamination/poisoning and/or model weight manipulation.
    c. Methods exploiting potential vulnerabilities in common AI development pipelines/frameworks.
    d. Methods for reliably executing manipulation in the open data/model ecosystem.
    e. Methods for executing a malicious middleware between a sensor and an AI-enabled system (either on device or in cloud application).
    f. Exploitations of application programming interface-based AI services to gain model information or to directly manipulate the model.
3. Electronic warfare (EW) effects for manipulating AI-enabled battlefield systems, to include:
    a. Methods of using EW for high precision sensor input modification over a variety of wavelengths, including electro-optical (EO) and other sensor modalities at a given distance to a sensor.
    b. Methods of jamming/manipulation EW-based communications with autonomous systems.
4. Physical manufacturing of adversarial effects, to include:
    a. Methods of automatic/rapid construction of “adversarial objects” from AI specification including 2D printing and 3D shapes.
    b. Materials research in 2D color printing material that has reduced glare and high-fidelity in print quality. These may include paper, cloth, or other printable material.
    c. Electronic displays that can adapt to ambient lighting condition to ensure a constant display from an EO sensor.
    d. Programmable fabrics that can dynamically display patterns.
    e. 3D shape and material construction methods that can have precise and controllable interactions with Light Detection and Ranging (LIDAR) and Synthetic-aperture radar (SAR) sensors.
5. Other effects
    a. The list above is not necessarily exhaustive. Submissions that do not fit a specific category above may still be relevant but require a CONOP, as described in the previous section, to support the relevance of the submission. To guide submissions, consider how your techniques and/or tools may

serve as an "attack vector/effect" for an autonomous AI-enabled system deployed by the DoD.

Responses to area 3, as well as area 4.a and 4.e, will be submitted as controlled unclassified information (CUI) at a minimum per the security instructions in this RFI or as a classified response if the content is known or suspected to be classified. Responses to other topic areas may require CUI or classified handling depending on content.

SUBMISSION INSTRUCTIONS:

Responses to this RFI should be submitted no later than, 5:00 p.m. Eastern Time on 28 February 2025.

Unclassified responses to this RFI should be submitted to DARPA-SN-25-28@darpa.mil. NO CLASSIFIED INFORMATION SHOULD BE SENT TO DARPA-SN-25-28@darpa.mil

Controlled Unclassified Information (CUI) responses:

For unclassified responses containing CUI, applicants will ensure personnel and information systems processing CUI security requirements are in place.

If an unclassified submission contains CUI or the suspicion of such, as defined by Executive Order 13556 and 32 CFR Part 2002, the information must be appropriately and conspicuously marked CUI in accordance with DoDI 5200.48. Each entity is responsible for assessing their technology and ensuring compliance with all export control laws and regulations. This RFI does not relieve any program participant from export control requirements.

Unclassified submissions containing CUI may be submitted via DARPA-SN-25-28@darpa.mil.

Entities submitting CUI responses must have an information system authorized to process CUI information IAW NIST SP 800-171 and DoDI 8582.01.

Classified responses:

For classified responses, applicants will ensure all industrial, personnel, and information systems processing security requirements are in place and at the appropriate level (e.g., Facility Clearance Level (FCL), Automated Information Security (AIS), Certification and Accreditation (C&A), and any Foreign Ownership Control and Influence (FOCI) issues are mitigated prior to submission. Additional information on these subjects can be found at https://www.dcsa.mil/.

If a submission contains Classified National Security Information or the suspicion of such, as defined by Executive Order 13526, the information must be appropriately and conspicuously marked with the proposed classification level and declassification date.

Submissions requiring DARPA to make a final classification determination shall be marked as follows:

"CLASSIFICATION DETERMINATION PENDING. Protect as though classified _____ *[insert the recommended classification level, e.g., Confidential, Secret, or Top Secret]*."

Entities choosing to submit classified information from other classified sources (i.e., sources other than DARPA) must ensure (1) they have permission from an authorized individual at the cognizant Government agency (e.g., Contracting Officer, Program Manager); (2) the submission is marked in accordance with the source Security

Classification Guide (SCG) from which the material is derived; and (3) the source SCG is submitted along with the submission.

Use transmission, classification, handling, and marking guidance provided by previously issued SCGs, the DoD Information Security Manual (DoDM 5200.01, Volumes 1 - 4), and the National Industrial Security Program Operating Manual, including the Supplement Revision 1 (DoD 5220.22-M and DoD 5200.22-M Sup. 1), when submitting Confidential, Secret, and/or Top Secret classified information.

Submissions containing both classified information and CUI must be appropriately and conspicuously marked with the proposed classification level, as well as ensuring CUI is marked in accordance with DoDI 5200.48.

Classified responses should be coordinated with DARPA prior to submission. Respondents wishing to provide a classified response should send an e-mail to DARPA-SN-25-28@darpa.mil with the subject line "Classified Coordination Requested." Respondents should allow at least three (3) business days for processing requests. NO CLASSIFIED INFORMATION SHOULD BE SENT TO DARPA-SN-25-28@darpa.mil.

To the maximum extent possible, respondents should submit non-proprietary information. If proprietary information is submitted, it must be appropriately and specifically marked. It is the respondent's responsibility to clearly define to the Government what is considered proprietary data. Any proprietary information should be clearly labeled as "Proprietary." DARPA will disclose submission contents only for the purpose of review by DARPA staff, other Government agencies, or DARPA Support Contractors.

NOTE: DARPA may conduct individual discussions with respondents as necessary to gain a full understanding of the technical response. DARPA may contact respondents individually via e-mail.

FORMAT INSTRUCTIONS:

No formal template will be required, however responses to the RFI should be concise. Respondents should submit a single integrated response addressing the relevant areas described above. DARPA will only review responses submitted as an unprotected Microsoft Word or PDF file.

DARPACONNECT:

Entities who have not worked with DARPA before are encouraged to learn more about DARPAConnect, an initiative established to facilitate collaboration between DARPA and potential performers. The DARPAConnect team offers customized support, resources, and guidance on how to prepare your ideas for high-impact conversations with DARPA program managers. Please visit DARPAConnect.us to access a digital hub of online resources, including a curriculum for self-paced learning, personalized support, and in-person and virtual events. In addition to the self-paced online materials, the DARPAConnect team is able to schedule one-on-one conversations to discuss your specific ideas, questions, and paths to DARPA. You can use the contact form at DARPAConnect.us or email the DARPAConnect team directly at darpaconnect@darpa.mil to request assistance.

ADMINISTRATIVE:

This announcement contains all information required to submit a response. No additional forms, kits, or other materials are needed. All administrative and technical questions should be directed to DARPA-SN-25-28@darpa.mil. Please refer to the Special Notice number (DARPA-SN-25-28) in all correspondence. This RFI is issued solely for information and program planning purposes and does not constitute a formal solicitation for proposals or proposal abstracts; any SOW sent will be disregarded. In accordance with FAR 15.201(e), responses to this notice are not offers and cannot be accepted by the Government to form a binding contract. Submission of a

response is strictly voluntary and is not required to propose to subsequent Announcements (if any) or Solicitations (if any) on this topic. DARPA will not provide reimbursement for costs incurred in responding to this RFI. Respondents are advised that DARPA is under no obligation to acknowledge receipt of the information received or provide feedback to respondents with respect to any information submitted under this RFI.