PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-mz8a-cg20
Comments Due: May 28, 2025 **Submission Type:** API

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0208
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Dreadnode

General Comment

See attached file(s)

Attachments

Dreadnode_RFI 2025 National AI Strategic Plan

May 29, 2025

Re: Request for Information on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan

Docket ID No. NSF-2025-OGC-0001. Submitted Electronically.

Introduction

To lead in AI over the next 3 to 5 years, the U.S. must treat quality dataset analysis and behavioral anomaly research as core priorities for national security and critical infrastructure resilience. Current benchmarks, evaluations, and test beds do not yet capture the scale, complexity, or risks posed by emerging AI systems. Targeted R&D investments are urgently needed to develop standards for dataset integrity, detect behavioral anomalies in models—both adversarial and autonomous—and support software test beds that reflect real-world operating conditions. These efforts will strengthen the foundations of AI reliability and enable secure, scalable deployment across high-stakes domains. Progress will require national standards, open-source infrastructure, and cross-sector consortia connecting government, industry, and academia.

Increase Investments in Data Science and Quality Dataset Analysis

Any advancement in AI relies heavily on the acquisition and maintenance of **quality training data.** At Dreadnode, our focus on data science reflects a core belief: the structure and provenance of data define a model's performance ceiling—especially when benchmarked against human behavior and interaction.

For datasets involved in LLM pre-training and fine-tuning, U.S. federal funding should support R&D that quantifies dataset contamination and related impacts to model reliability. Examples of recommended research areas include:

- Quantifiable benchmarks for **dataset contamination, labeling consistency,** and **hidden bias detection** that correlate with specific AI failure modes including hallucination rates and capability degradation—with evaluation methods that work both through **direct dataset analysis** and **output-based behavioral testing**.
- Tools for **auditing training** and **fine-tuning corpora**, with metadata tagging, integrity verification, and systematic model-model performance comparisons.
- Evaluation of **adversarial data insertion** including boundary bypass triggers, backdoor activation patterns, and multi-stage poisoning attacks across different training phases.

Federal R&D should recognize the fundamental difference between open-source and proprietary training approaches. Open-source datasets enable direct contamination analysis and bias auditing, while proprietary datasets require **output-based evaluation methods** that can assess model quality without accessing training data. Investment should prioritize developing **black-box evaluation techniques** that measure dataset quality issues through model behavior analysis, ensuring consistent standards across both development paradigms without compromising competitive advantages.

Post-deployment behavioral and environmental data collection requires the same level of federal investment, with an emphasis on **robust telemetry and observability**. These R&D efforts can span the following:

- **Telemetry instrumentation standards** that track input-output mappings, intermediate agent actions, reasoning processes, multi-agent communications, resource consumption patterns, and confidence scoring across distributed AI systems.
- Real-time **LLM instrumentation tools** that integrate cybersecurity observability frameworks (SIEM, SOAR, XDR) with AI-specific anomaly detection, normal behavior patterns, and automated threat response capabilities designed for production-scale deployments.

Finally, as agent-based AI systems evolve over time, it's imperative that we better understand cumulative agent behavior. This area of focus should include pre- and post-task reasoning traces to enable root cause analysis of behavioral divergence. Investment areas include:

- **Temporal behavior logging**, including memory updates, world model shifts, and sequential decision-making.
- **Attack replay datasets** for repeated benchmarking of known exploits or manipulative interactions across foundation models.

Federal investments can also facilitate a **National AI Data Quality Assurance Consortium** between interested public and private sector stakeholders, building on existing efforts like open-source data contamination knowledge shares and DARPA programs such as SEMAFOR. This consortium would focus on:

- Establishing benchmarks for AI systems seeking integration into government and critical infrastructure environments.
- Standardizing metrics for dataset integrity, contamination detection, and bias assessment that determine whether AI models meet federal deployment standards.

- Leveraging federal threat intelligence, industry deployment experience, and academic methodology to inform benchmark standards.

Stakeholders across self-regulated proprietary frontier model companies and open-source large language model developers would help shape the qualification criteria that ultimately ensure only rigorously vetted AI systems are permitted to support sensitive government operations.

Promote Focused Research around AI Behavioral Anomalies

Dreadnode focuses on offensive AI to rigorously test model boundaries and expose exploitable behavior. We prioritize red teaming and cyber evaluations because securing the AI-enabled future requires deep experimentation and a clear understanding of system failure modes. Federal investment is needed in adversarial testing infrastructure or red teaming, and these evaluations must:

- Include benchmarks for **tool exploitation**, simulating how agents interact with external APIs, file systems, or shell environments.
- Detect **goal obfuscation**, where a model conceals its capabilities or objectives under scrutiny.
- Evaluate **adversarial goal hacking**, where an agent is prompted or tricked into reshaping its utility function or rules of engagement.
- Test for **agentic manipulation**, including deceptive, strategic, or emergent behaviors over time and across tasks.

Non-adversarial or embedded behavioral anomalies are also highly prevalent, as they speak to compute limitations or faulty system dynamics. Some examples include:

- **AI sandbagging**, which references a model's intentional underperformance or capability suppression.
- **Alignment faking**, where a model appears to be aligned with human instruction, but is behaving with deceptive intent.
- **Resource-constrained evaluation**, which measures AI performance under operational and environmental stress, to include **fallback behavior** and **graceful degradation**.

The TEST AI Act of 2025 proposes a NIST-led pilot program that uses testbeds to develop measurement standards for the evaluation of artificial intelligence systems, and for other purposes. Should Congress pass this act into legislation, U.S. federal funding should support software testbeds that simulate the following conditions:

- **Air-gapped, degraded compute**, or **manipulated toolchain** conditions.
- **Network latency, throttling**, or **packet injection** scenarios.
- **Evaluation-mode** and **production-mode** conditions, to assess AI behavioral deception, discrepancies, and drift across both environments.

Robust software testbeds will not only assist with evaluations of AI model behavior, but they can also enable the testing and optimization of beta software releases from government programs. One such example includes the Cyber Reasoning Systems designed to find and fix open-source vulnerabilities at scale following the AI Cyber Challenge Final Competition in August 2025.

Beyond individual testbed capabilities, coordinated evaluation efforts require systematic collaboration across sectors. The U.S. government can facilitate an **AI-Enabled Red Teaming Consortium** to leverage these software testbeds, building on the work performed by companies like Dreadnode, federally funded research and development centers (FFRDCs) like MITRE CALDERA, and industry-led expertise. A public-private partnership of this nature would inherently improve AI operational resilience by:

- Stress testing models across a **robust attack surface**.
- Accelerating **automated vulnerability discovery and remediation efforts**.
- Enhancing defense development through **real-time intelligence sharing** and **collaborative countermeasure strategies**.

Any related findings can inform **adversarial capability and autonomous behavioral drift** benchmarks and evaluation mechanisms.

Conclusion

The next 3-5 years require focused investment in dataset integrity research, behavioral anomaly detection, and adversarial testbed infrastructure to ensure U.S. AI capabilities remain resilient against nation-state threats. Success will be measured by standardized auditing tools, quantifiable behavioral benchmarks, and operational testbeds that enable confident AI deployment across civilian and defense applications. We strongly encourage OSTP and its federal partners to prioritize these investments as national security imperatives in the 2025 National AI R&D Strategic Plan.

Contact

To get in touch with the team for further discussion.