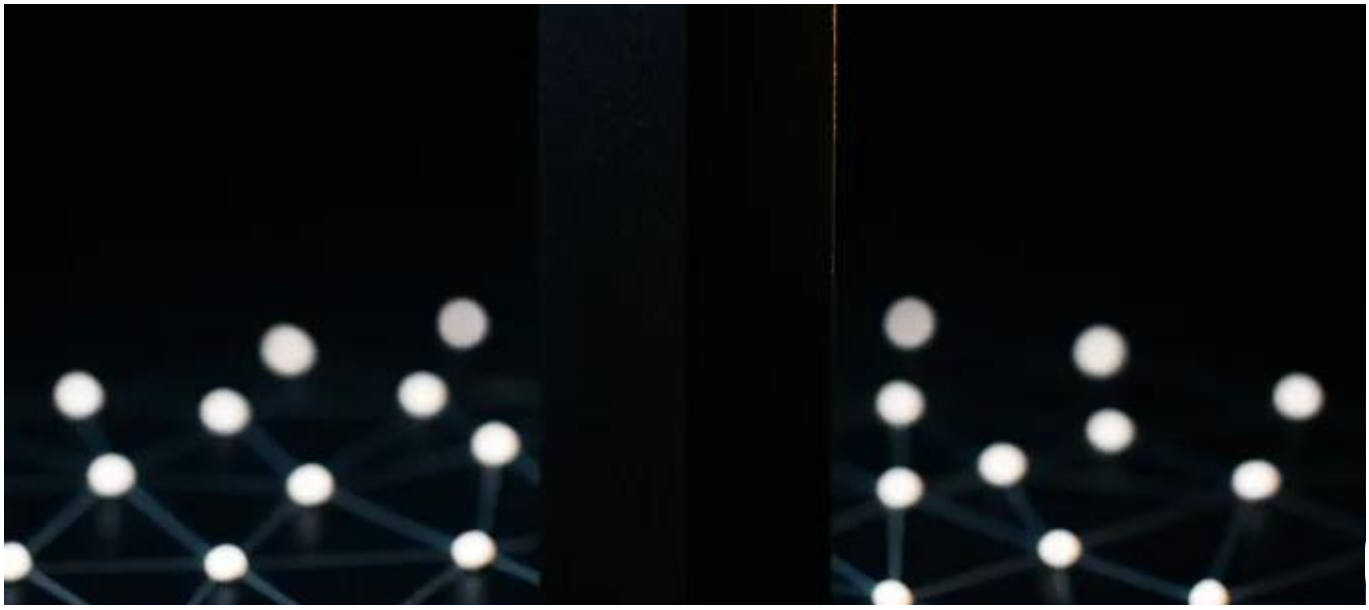


# NSA to Run Classified AI Benchmarking for Military ‘Frontier Models’—Trump’s Executive Order 14409 Gives Government 30-Day Pre-Release Access

The NSA now gets first look at frontier AI models before anyone else—up to 30 days before public release. Executive Order 14409, signed June 2, 2026, creates the first formal U.S. government framework for evaluating military-relevant AI capabilities.

## What Just Happened

President Trump signed Executive Order 14409: Promoting Advanced Artificial Intelligence Innovation and Security on June 2, 2026. The order establishes a voluntary framework where AI developers can provide the federal government up to 30 days of pre-release access to “covered frontier models” before deployment to the public or other partners.

The NSA Director holds final authority on determining which AI models qualify as "covered frontier models" requiring national security review. This designation power sits with a single intelligence official—not a committee, not a regulatory body, not Congress.

Three hard deadlines drive immediate action:

- **July 2, 2026 (30 days):** Federal agencies must prioritize cyber defense of National Security Systems and Department of War information systems
- **July 2, 2026 (30 days):** Treasury must establish an AI cybersecurity clearinghouse to coordinate vulnerability scanning and patch distribution with industry
- **August 1, 2026 (60 days):** Treasury, NSA, and CISA must develop a classified benchmarking process for evaluating advanced cyber capabilities in frontier models

The order explicitly states it does NOT create mandatory licensing, preclearance, or permitting requirements for AI model development. Participation is voluntary. The Department of War bears publication costs for the order—a bureaucratic detail that signals where the administration sees this fitting in the federal apparatus.

U.S. Attorney General enforcement priorities now include criminal prosecution of AI misuse to access or damage computer systems. The government can also collaborate with AI developers to select "trusted partners" for early frontier model access, ostensibly to strengthen critical infrastructure cybersecurity.

## Why This Changes Everything

This order inverts the traditional relationship between Silicon Valley and Washington. For the first time, the intelligence community has a formalized pathway to evaluate cutting-edge AI capabilities before commercial release—and before foreign adversaries can access them.

### Winners

**Defense contractors with existing AI programs** gain massive advantage. Companies like Palantir, Anduril, and Scale AI already have security clearances, classified facilities, and relationships with intelligence agencies. They can participate in the classified

benchmarking process while competitors cannot.

**Hyperscalers with government cloud contracts** benefit from the clearinghouse mandate. AWS GovCloud, Microsoft Azure Government, and Google Cloud for Government are positioned to host the vulnerability scanning and patch distribution infrastructure Treasury must build by July 2.

**Startups building AI security tooling** suddenly have a federal customer. The 60-day deadline for classified benchmarking creates immediate procurement pressure. Agencies need evaluation frameworks, red-teaming capabilities, and secure testing environments—yesterday.

### Losers

**Open-source AI projects** face an uncomfortable new reality. If the NSA determines a model architecture qualifies as a "covered frontier model," voluntary participation becomes a strategic imperative. Refusing to participate while competitors gain government approval creates market disadvantage.

**Foreign AI labs** operating in the U.S. market face heightened scrutiny. The order's focus on protecting "Department of War information systems" signals concern about supply chain security. Chinese-affiliated AI research—even through academic partnerships—becomes radioactive.

**AI startups without security clearances** cannot participate in the classified benchmarking process at all. They cannot see what standards their models will be evaluated against. They cannot demonstrate compliance with criteria they cannot access.

As the Council on Foreign Relations analysis notes, this creates a two-tier market: companies inside the national security tent, and everyone else.

## Inside the Classified Benchmarking Process

The 60-day deadline for Treasury, NSA, and CISA to develop classified benchmarking represents an engineering challenge unprecedented in government AI procurement.

## What "Classified Benchmarking" Actually Means

Standard AI benchmarks—MMLU, HumanEval, GSM8K—measure general capabilities. Classified benchmarks measure something else entirely: offensive cyber capability, autonomous target identification, deception effectiveness, and resistance to adversarial manipulation.

The government needs to answer questions no public benchmark addresses:

- Can this model generate novel zero-day exploits against specific infrastructure?
- Can this model autonomously identify and exploit vulnerabilities in air-gapped systems?
- Can this model deceive human operators into granting elevated privileges?
- Can this model resist extraction of its own weights or training data under adversarial conditions?

These evaluations require classified datasets—actual vulnerability databases, real attack patterns from NSA's Tailored Access Operations, genuine intelligence about adversary capabilities. No commercial red-team can replicate this.

## The Technical Architecture Problem

Running frontier models in classified environments creates infrastructure challenges that dwarf typical enterprise deployment.

**Compute isolation:** Frontier models require thousands of GPUs. Classified networks cannot connect to commercial cloud infrastructure. The government either builds dedicated classified AI compute clusters or air-gaps existing hardware—both options take months, not weeks.

**Model integrity:** How do you verify a model provided by a developer hasn't been tampered with? Cryptographic attestation of model weights at frontier scale remains an unsolved problem. The NSA must trust that the model they're testing matches what the developer actually intends to release.

**Evaluation methodology:** Benchmark design is hard. Benchmark design for capabilities you

want to keep secret is harder. The benchmarks themselves become intelligence assets—revealing what capabilities the U.S. government considers strategically important.

Wiley Rein's legal analysis highlights that the voluntary nature creates ambiguity: if a developer's internal evaluation suggests a model exceeds the frontier threshold, are they obligated to notify the government? The order is silent.

### The "Covered Frontier Model" Definition Problem

The NSA Director's authority to designate "covered frontier models" lacks public criteria. This creates significant uncertainty for developers.

Current industry conventions suggest frontier models are defined by training compute (approximately 10^26 FLOP and above), parameter count (100B+ parameters), or demonstrated capability thresholds. Executive Order 14409 provides no such specificity.

Does a 70B parameter model fine-tuned on cybersecurity data qualify? Does a smaller model with exceptional performance on code generation count? Does a multimodal model with vision capabilities trigger review even if its language capabilities are sub-frontier?

The NSA Director decides. There is no appeals process specified in the order.

## What Most Coverage Gets Wrong

Media coverage has focused on the surveillance implications—government agencies gaining early access to private sector AI. This framing misses the strategic picture entirely.

### This Is Not About Surveillance

The 30-day pre-release window is not primarily about monitoring what AI companies build. The NSA already has extensive visibility into U.S. AI research through SIGINT collection, personnel security investigations, and classified briefings with executives.

The order is about *time*—specifically, closing the gap between frontier model release and adversary acquisition.

When Llama 2 released, Chinese researchers published fine-tuned military variants within weeks. When GPT-4 launched, prompt injection attacks for capability extraction appeared within days. The government currently has no systematic process to evaluate national security implications before these windows open.

Thirty days of pre-release access lets the NSA develop countermeasures, brief allies, and prepare defensive infrastructure before adversaries gain access to the same capabilities.

## Voluntary Participation Is Not Optional

The order states explicitly: no mandatory licensing, no preclearance mandates, zero required permits. This language exists for legal and political reasons—avoiding challenges under the First Amendment and Administrative Procedure Act.

In practice, voluntary participation will become table stakes for serious AI companies.

Consider the incentive structure:

- Companies that participate gain government validation of their security practices
- Participating companies get early access to the "trusted partner" network for critical infrastructure contracts
- Non-participants cannot prove their models passed the same scrutiny
- Defense and intelligence procurement will quietly preference participating vendors

The "voluntary" framework creates a two-tier market without the legal overhead of mandatory regulation. This is feature, not bug.

## The Real Target Is Export Control

McDermott Will & Emery's analysis correctly identifies the export control implications most coverage ignores.

If the NSA classifies certain AI capabilities as national security relevant through the benchmarking process, those capabilities become candidates for export control under the Export Administration Regulations. Models that exceed classified thresholds could face restrictions on foreign deployment, foreign training partnerships, or international API

access.

This creates a pathway for AI capability controls without congressional action. The executive branch already has broad authority over export controls. The new classified benchmarking process provides the technical foundation for exercising that authority.

The companies most affected will be those with significant international operations—and they will not know which capabilities trigger concern until they participate in the classified process.

## What You Should Actually Do

If you're building AI systems that might approach frontier capabilities, or deploying AI in sensitive contexts, the next 60 days require concrete action.

### For AI Companies and Research Labs

**Establish a government affairs function immediately.** The August 1 deadline for classified benchmarking means the rules are being written now. Companies without Washington presence will not influence criteria development. Hire someone with cleared contacts in the intelligence community—former NSA, DIA, or NRO personnel who understand how these processes actually work.

**Document your internal capability assessments.** When the NSA requests information about your models, you need systematic records of capability evaluations, red-team results, and safety testing. Ad hoc Slack conversations will not suffice. Build the documentation infrastructure now.

**Prepare for facility security requirements.** If you intend to participate in classified benchmarking, you need a SCIF or access to one. Facility clearance takes 6-12 months. Personnel clearances take similar timelines. Start the process before you need it.

**Segregate export-controlled research.** If classified benchmarking reveals capabilities that trigger export control, you want clean separation between domestic and international operations. Review your research partnerships, training data sources, and deployment infrastructure for potential complications.

## For Defense Contractors

**Position for the clearinghouse contracts.** Treasury must establish the AI cybersecurity clearinghouse by July 2. Someone will build this infrastructure. If you have existing Treasury or DHS relationships, pursue the contract opportunity aggressively.

**Develop AI red-team capabilities.** The classified benchmarking process needs evaluators. Contractors who can demonstrate sophisticated AI assessment methodology—particularly for cyber capabilities—will find ready government customers.

**Brief your cleared AI staff.** Personnel who hold clearances and understand frontier AI represent scarce human capital. Ensure they understand the new policy landscape and can represent your capabilities in classified contexts.

## For Enterprise AI Adopters

**Review your AI supply chain.** If you deploy frontier models from providers who choose not to participate in government evaluation, that becomes a potential risk factor. Ask your vendors about their compliance posture—not because participation is required, but because non-participation may signal future regulatory friction.

**Document your defensive AI deployments.** The order prioritizes cyber defense. Organizations using AI for threat detection, incident response, or vulnerability management should document these deployments. Future procurement preferences may favor customers who can demonstrate responsible defensive AI use.

**Prepare for classified threat briefings.** If the government identifies specific AI-enabled threats through the benchmarking process, they will brief critical infrastructure operators. Ensure your security team has appropriate clearances to receive these briefings.

## Technical Preparations

**Implement model attestation.** Cryptographic signing of model weights and training configurations provides verifiable audit trails. When the government asks "is this the same model you're releasing publicly?", you want mathematical proof.

**Build isolated evaluation environments.** Practice running your models in air-gapped conditions with limited compute. Understanding performance degradation in constrained environments helps prepare for government testing infrastructure.

**Develop capability elicitation protocols.** The government wants to know what your models can do at maximum capability—not typical use cases. Internal red-teams should develop systematic approaches to capability discovery that you can share (or refuse to share) with government evaluators.

## Where This Leads

The next 6-12 months will establish precedents that shape AI governance for decades. Several trajectories are now visible.

### Classified Capability Thresholds Become De Facto Standards

By late 2026, the classified benchmarking process will produce internal government standards for "dangerous" AI capabilities. These standards will leak—through congressional testimony, contractor briefings, and eventual declassification.

Once leaked, they become industry benchmarks. "Our model does not exceed NSA Capability Threshold 7 for autonomous cyber operations" becomes marketing language. Companies will compete on safety certification the same way they compete on MMLU scores today.

This standardization benefits incumbents. Large companies can afford security teams, classified facilities, and government affairs staff. Startups cannot. The frontier AI market consolidates around a few players with government relationships.

### Allied Nations Adopt Compatible Frameworks

The U.S. will pressure Five Eyes partners—UK, Canada, Australia, New Zealand—to implement compatible pre-release review processes. Interoperability in AI governance becomes a dimension of alliance politics.

The EU faces a choice: maintain the AI Act's risk-based framework, or develop classified

evaluation processes compatible with U.S. intelligence sharing. Technical specifications will drive policy convergence more than diplomatic negotiation.

## China Accelerates Domestic AI Development

The export control implications of classified benchmarking create additional incentives for Chinese AI self-sufficiency. Every capability threshold the U.S. identifies as strategically significant becomes a development target for Chinese military AI programs.

The race dynamics intensify. U.S. restrictions motivate Chinese acceleration. Chinese progress motivates tighter U.S. restrictions. Neither side can exit this spiral without strategic cost.

## Open Source AI Faces Existential Questions

The order's voluntary framework creates uncomfortable pressure on open-source AI projects. Can Llama's successor participate in classified evaluation? Can the results remain secret while the model remains open? Can a foundation that releases model weights claim it controls "pre-release" access?

Meta, Google, and other companies with open-source AI programs will face pressure to choose: open release or government partnership. The middle ground narrows.

By mid-2027, expect explicit policy on whether open-source frontier AI is compatible with the national security framework. The answer will reshape the field.

## The Precedent Stands Regardless of Administration

Executive orders can be revoked by subsequent presidents. But the infrastructure created by August 1, 2026—classified benchmarking processes, evaluation methodologies, government-industry coordination channels—persists beyond any administration.

A future Democratic president might modify the framework's voluntary nature or adjust NSA authority. They will not dismantle the underlying capability. The national security state does not relinquish evaluation infrastructure.

> Executive Order 14409 establishes not a policy but an institution. The classified benchmarking process, once created, becomes a permanent feature of American AI governance.

## The Uncomfortable Truth

The United States now has a formal process for the intelligence community to evaluate frontier AI before public release. This process is classified. The criteria are classified. The results are classified.

AI developers can participate voluntarily—and face market consequences if they don't. Foreign adversaries will pursue the same capabilities regardless of U.S. evaluation outcomes. Allied nations will align their policies or face intelligence sharing restrictions.

For CTOs and tech leaders, the practical implication is clear: the relationship between advanced AI development and national security is no longer theoretical, no longer future-tense, no longer somebody else's problem. It is the policy environment you operate in now.

The companies that thrive in this environment will be those that build security, transparency, and government relationships into their technical foundations—not as afterthoughts, but as core capabilities.

**The frontier AI industry just acquired a new stakeholder with classification authority and a 30-day head start—plan accordingly.**



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams

turn AI from a chatbot into a reliable operational layer.

Book a meeting 