MODERNIZING GLOBAL VULNERABILITY STANDARDS

Reforming our Standard, Norms, and Infrastructure to Meet the AI Vulnerability Challenge

A Public Policy Position on Verification, Access, Disclosure, and Institutional Accountability

Corey Thomas, Executive Chair, Rapid7

Sabeen Malik, Chief Global Policy Officer, Rapid7

Notes on Evidence and Methodology

The arguments in this paper rely primarily on capability claims published by Anthropic, OpenAI, and Google DeepMind. These are vendor-published benchmarks that have not been independently verified. We treat them as well-supported hypotheses motivating precautionary action, not settled findings.

The infrastructure strain documented by NIST and the Zero Day Clock data on time-to-exploitation are independently corroborated and carry greater evidentiary weight.

Rapid7 was accepted into OpenAI's Trusted Access for Cyber program and remains under consideration at Anthropic.

EXECUTIVE SUMMARY

AI has invalidated the assumptions behind every major vulnerability prioritization tool in use today. CVSS, the CISA Known Exploited Vulnerabilities catalog, the Exploit Prediction Scoring System, and the National Vulnerability Database were all designed for human-speed discovery, manageable volumes, and exploitability confirmed after the fact. None of those conditions still hold. This paper identifies where each tool breaks and what must be done to fix them.

In April 2026, Anthropic, OpenAI, and Google DeepMind each announced production-grade AI systems capable of discovering, chaining, and in some cases remediating software vulnerabilities at machine speed. The Stanford HAI AI Index 2026 Cybench benchmark documents that unguided AI agents solve rates on cybersecurity tasks rose from 15% to 93% in a single year. These are deployed capabilities on a steep improvement curve, not research experiments.

Yet the institutional infrastructure surrounding vulnerability management (the databases, scoring frameworks, disclosure timelines, and government programs that every organization depends on) was built for human-speed discovery. It is already strained beyond capacity. CVE submissions grew 263% between 2020 and 2025 from human-speed growth alone. NIST acknowledged in April 2026 that the National Vulnerability Database can no longer keep pace, shifting to risk-based triage. The prioritization tools that organizations use to decide what to fix first each break in identifiable ways under AI-era conditions.

This paper argues that the prioritization gap is the most urgent and least addressed dimension of the problem. It proposes two categories of reform:

Policy reforms requiring government action: update the Vulnerabilities Equities Process for AI-scale discovery; rebuild CVE/NVD infrastructure for AI-era volumes; mandate capability disclosure packages; establish binding international coordination frameworks, and confirm CISA leadership.

Access and verification standards for the security community: independent verification before access expansion; structured, curated access through transparent processes; and rigorous data standards for published capability claims.

The frontier model providers building these capabilities deserve credit for acting responsibly as they invent policy in real time. But individual programs cannot substitute for common standards backed by independent verification, institutional accountability, and international coordination. The window to build those standards is short and closing.

I. THE THRESHOLD

AI-driven vulnerability discovery has crossed a threshold the security community anticipated for years but is not prepared for. In April 2026, Anthropic announced that its Claude Mythos Preview model had autonomously discovered thousands of zero-day vulnerabilities across major operating systems, browsers, and open-source components. OpenAI scaled its Trusted Access for Cyber program to thousands of verified defenders. Google DeepMind launched CodeMender, an AI agent that finds and fixes critical code vulnerabilities. The convergence of production capability claims across three independent organizations is significant even before independent verification is complete.

To understand why standards did not already exist for this moment, consider where these tools came from. Large language models were developed by AI research organizations, not security companies or defense contractors or regulated industries. The intellectual tradition is academic machine learning: publish findings, share models, iterate openly. Those norms run counter to the pre-certification mindset that governs security products. When a security company builds an intrusion detection product, it enters a world of procurement frameworks, certification expectations, and enterprise sales cycles that impose informal standards before regulators act. When an AI lab discovers that its general-purpose model can find zero-day vulnerabilities, it has stumbled on an emergent capability of something built for other reasons entirely. There was no obvious moment to impose standards because the capability was not designed. It arrived.

That gap is now a liability. The security industry spent two decades building processes for a world where vulnerability discovery would remain at human speed. Coordinated disclosure timelines assume researchers need weeks to find and verify a flaw. The National Vulnerability Database assumes human-speed enrichment can keep pace with submissions. Prioritization frameworks assume the volume of critical findings is manageable with current staffing. Every one of those assumptions is now wrong.

Bruce Schneier and others have observed that finding vulnerabilities for the purpose of fixing them is structurally easier for AI than finding and exploiting them, and that defenders may hold a durable advantage because remediation can scale faster than weaponization. That asymmetry, if real and lasting, is the best possible outcome. But it has not been empirically established, and it does not reduce the urgency of the infrastructure problem. Even if defenders hold a permanent advantage in what AI can do, that advantage is worthless if the scoring frameworks, databases, disclosure timelines, and government programs around them cannot absorb and act on what AI produces.

Whether AI vulnerability discovery is good or bad for defenders depends entirely on whether the institutional infrastructure gets rebuilt fast enough to use these tools for defense before adversaries use equivalent tools to cause serious damage. Open-weight models are behind today, but the gap will narrow. Within months, the ability to discover and chain software vulnerabilities at machine speed will be available to a much broader set of actors, including adversaries. The technology moved a decade ahead in a single year. Everything else needs to catch up.

This paper argues for action on two fronts: institutional policy reforms that no industry program can substitute for, and access and verification standards governing how AI vulnerability discovery capabilities are distributed and validated. These are not independent problems, and their correct sequencing matters.

II. WHAT HAS BEEN DEMONSTRATED

Anthropic's Claude Mythos Preview, per Anthropic's published reports, autonomously discovered thousands of zero-day vulnerabilities including flaws that had survived decades of human review. Anthropic reported an 83% reproduction rate on known vulnerabilities and demonstrated the ability to chain multiple Linux kernel vulnerabilities into a full privilege escalation without human guidance. These claims have not been independently verified. They are cited here as motivating evidence, not settled benchmarks.

OpenAI developed GPT-5.4-Cyber for defensive security work and, through its Codex Security initiative, contributed to over 3,000 critical and high-severity vulnerability fixes. It committed $10 million in API credits to accelerate cyber defense. Google DeepMind's CodeMender has provided 72 security fixes to open-source projects, all human-reviewed before submission. Google also updated its Frontier Safety Framework to add capability tracking for security-relevant AI functions.

Each laboratory has taken a different approach to distribution, and each deserves recognition for building responsible programs in the absence of established standards. Anthropic's Project Glasswing is a coalition of organizations, including major technology companies, financial institutions, and open-source foundations. The program is focused on active defensive security work in which participating organizations use Claude Mythos Preview to find and fix vulnerabilities in real-world critical software infrastructure, and partners within the initiative are able to deploy it in their security workflows, including against production systems. OpenAI's Trusted Access for Cyber program has a structured application process, responds within a week, and has scaled to thousands of defenders across multiple sectors. Google's CodeMender focuses on direct remediation of open-source vulnerabilities rather than broad tool access.

Each model reflects different assumptions about the balance between control and speed. The absence of common standards means every organization seeking access must navigate a different process with different criteria and different levels of transparency. That is not sustainable when the threat these capabilities address affects every organization that runs software.

III. RETOOL THE PRIORITIZATION INFRASTRUCTURE STACK

The previous section describes what AI can now find. This section addresses a different and arguably more urgent problem: even when vulnerabilities are found and processed, the tools organizations use to evaluate them were built for a threat environment that no longer exists.

The CVE identifier system, the CVSS severity scoring framework, CISA's Known Exploited Vulnerabilities catalog, and the Exploit Prediction Scoring System all rest on the same foundational assumption: exploitability is confirmed after the fact, and each vulnerability is assessed on its own. AI breaks both. Once independent verification establishes that AI-demonstrated exploitability is reliable, the prioritization infrastructure must be updated to incorporate it.

Where Each Tool Breaks

CVSS grades each vulnerability individually, with no mechanism to flag that two medium-severity flaws, harmless in isolation, can be combined into a path that gives an attacker full control of a system. Mythos demonstrated exactly this by chaining multiple Linux kernel vulnerabilities into a full privilege escalation. No individual CVSS score would have flagged any component of that chain as urgent. An organization following standard practice would have deprioritized every piece of it.

The CISA Known Exploited Vulnerabilities (KEV) catalog asks a narrower question: has this vulnerability been actively exploited in the wild? CISA (the Cybersecurity and Infrastructure Security Agency, the lead federal agency responsible for coordinating vulnerability disclosure and cybersecurity guidance across the civilian government) maintains KEV as a binding directive (BOD 22-01) requiring every Federal Civilian Executive Branch agency to remediate listed vulnerabilities on fixed timelines. It is the single most influential prioritization signal in both government and enterprise security. But KEV is retrospective and binary. A vulnerability must already have been weaponized in the wild before it qualifies for entry. When mean time-to-exploitation has fallen below one day, organizations are often remediating vulnerabilities already being used against them before the catalog entry is created. AI-demonstrated exploitability, meaning a working exploit generated before any human attacker has found the flaw, has no entry path into KEV under current rules.

The value of the Known Exploited Vulnerabilities catalog is precision. Its authority derives not from size but from signal quality: every entry represents a confirmed, actionable threat that defenders must address. That signal cannot survive AI-scale volume without structural control. We propose that AI-demonstrated exploitability qualify for KEV entry only when it meets a higher evidentiary bar than current in-the-wild confirmation requires: independent reproduction by a verified third party, reachability confirmed in at least one production environment, and full chain documentation where multiple vulnerabilities compose the exploit path. This is not a higher bar for its own sake. It is the bar required to

keep KEV trustworthy when the volume of candidate findings grows by an order of magnitude. A KEV entry that defenders cannot act on with confidence is worse than no entry at all. The goal is not a larger catalog. It is a catalog that remains the single most reliable prioritization signal in both government and enterprise security, even as the threat environment it tracks accelerates beyond human speed

EPSS (the Exploit Prediction Scoring System) predicts the probability that a given vulnerability will be exploited within 30 days, using a model trained on historical attacker behavior. Two problems emerge under AI pressure. First, historical patterns may not reflect what AI-assisted attackers can now do, meaning the model may systematically underpredict risk for an entirely new capability class. Second, EPSS cannot score vulnerabilities that have not been publicly disclosed. Flaws discovered by AI and patched before any public announcement generate no data for the model to learn from. The better defensive AI discovery works, the less signal EPSS has to calibrate against. This is a structural feedback problem built into the current architecture.

Three Reforms

First: Update BOD 22-01 to recognize AI-demonstrated exploitability. CISA should add AI-demonstrated exploitability as a formal KEV entry criterion alongside confirmed in-the-wild exploitation. Because BOD 22-01 makes KEV binding on every Federal Civilian Executive Branch agency, changing what qualifies for entry rewrites what the federal government must fix and by when. That signal ripples into the private sector because most major organizations treat KEV as a prioritization input regardless of whether the directive applies to them directly. Three conditions must be met: a definition of what constitutes a verified AI exploit demonstration; a disclosure process that preserves the vendor notification window before a KEV entry is created; and updated remediation timelines for the new category, which may warrant a different window than the existing two-week mandate. This reform depends directly on independent verification and cannot be implemented until AI exploit demonstrations have been established as reliable.

Second: Add chaining-risk metadata to NVD entries. NIST should be directed to add a structured metadata field to CVE records flagging whether a vulnerability is a known component of an exploit chain, the severity of the end-state exploit, and what the other required components are. CVSS will not incorporate combinatorial risk quickly because it is a standards process on a multi-year cycle, but NIST does not need to wait. This gives organizations a chain-risk signal through existing NVD infrastructure without requiring CVSS to change. It also creates a dataset that EPSS and other prediction tools can incorporate into their training over time, gradually improving the prediction layer without requiring a full rebuild. Congress should fund this as a specific, mandatory deliverable within the NVD redesign, not a discretionary enhancement that gets deferred when budgets tighten.

Third: Require reachability guidance alongside discovery findings. AI discovery programs operating under verified access should be required to produce reachability guidance alongside their findings. A high CVSS score tells you how severe a vulnerability is in the abstract. It does not tell you whether that vulnerability can actually be triggered in your environment. A flaw in a library function that no deployed application ever calls is not a real threat, regardless of how it scores. Reachability analysis closes that gap by telling

you whether the vulnerable code path is live given how your organization has deployed and configured its software. Requiring discovery programs to include standardized reachability metadata (whether exploitation depends on particular configurations, network exposure, or call paths) distributes that intelligence to every recipient rather than reserving it for the few with capacity to reproduce the work. CISA should develop the metadata schema, and mandatory capability disclosure packages should require laboratories to include it as a condition of verified access.

These three reforms together shift the prioritization infrastructure from telling organizations how severe a vulnerability is in theory to telling them how dangerous it is in practice, in their actual environment. That shift is overdue regardless of AI. AI just made it impossible to defer any longer.

IV. GOVERNMENT REFORMS FOR AI VULNERABILITIES

The prioritization reforms above address how findings are evaluated. The following five reforms address the institutional and regulatory infrastructure that must change to absorb AI-era vulnerability discovery at scale. No industry program can fill these gaps.

1. Update the Vulnerabilities Equities Process

The Vulnerabilities Equities Process (VEP) is the government's framework for deciding whether to disclose or retain discovered vulnerabilities. It operates under a 2017 charter designed for individual vulnerabilities discovered through human analysis. It was not designed for a world in which an AI system produces thousands of zero-day findings in a single run, each requiring a separate equities determination, some involving private-sector actors operating under government access programs.

Three AI-specific questions the current VEP does not address must be resolved: whether the method of discovery changes the equities analysis; how the government handles batch disclosures at AI scale rather than case-by-case; and how VEP obligations attach when AI capabilities are held by private-sector partners rather than government agencies. ONCD should convene an expedited reform working group with CISA, NSA, DoD, and the intelligence community to produce an updated charter.

2. Require Mandatory Capability Disclosure Packages

AI laboratories making material capability claims about vulnerability discovery should be required, as a condition of federal procurement and where feasible through regulation, to publish standardized disclosure packages: benchmark composition and selection methodology, operational parameters, false positive rates, failure mode documentation, and performance variation across target types outside the training distribution. The Foundation Model Transparency Index score for frontier labs dropped from 58 to 40 in 2025, meaning the field is actively moving away from disclosure as capabilities accelerate. A mandatory disclosure requirement creates the legal and procurement basis for reversing that trend.

Disclosure packages should specifically require laboratories to report on the defender-attacker asymmetry: whether the model's ability to generate working exploits is advancing at the same rate as discovery, and what evidence supports any claim that defensive use structurally outpaces offensive use. This is data that policymakers need and currently cannot obtain.

3. Rebuild CVE and NVD Infrastructure for AI-Era Volumes

CVE submissions increased 263% between 2020 and 2025, driven by human-speed discovery growth, not yet by AI-scale production. NIST's April 2026 shift to risk-based prioritization is a structural admission that the current system cannot absorb what it

already faces. AI-driven discovery will accelerate volume by an order of magnitude. The 90-day coordinated disclosure window was calibrated for a world where exploitation took weeks; the Zero Day Clock now shows mean time-to-exploitation below one day.

Congress should fund a redesigned enrichment architecture: ML-triage before human review, distributed CVE Numbering Authority responsibilities across allied bodies including ENISA, and disclosure timelines that distinguish between a human researcher finding a flaw over months and an AI system finding thousands in a single run. The Advanced Artificial Intelligence Security Readiness Act and the FY 2026 NDAA AI cybersecurity governance provisions are steps in the right direction, but neither addresses disclosure standards for AI-discovered vulnerabilities specifically. That gap requires a dedicated mandate.

4. Establish Binding International Coordination Frameworks

The EU Cyber Resilience Act's mandatory vulnerability reporting obligations take effect on 11 September 2026. The EU AI Act's Article 73 incident reporting framework becomes effective on 2 August 2026. The US has no equivalent binding requirements. The UK NCSC, the International AI Safety Report, and secure AI development guidance co-authored by 18 countries have all reached the same threat assessment. US domestic policy has not matched it with a binding response framework.

CVE is a US-anchored program in a globally distributed threat environment. Multinational organizations navigating inconsistent frameworks will route compliance obligations toward the binding standard, which is increasingly European. The US should engage bilateral and multilateral partners through the Quad, Five Eyes, and the US-EU Trade and Technology Council to establish shared standards for AI vulnerability access programs, coordinated disclosure timelines calibrated to AI-velocity discovery, and mutual recognition of CVE Numbering Authority decisions. The goal is not a single global governance body but a minimum set of interoperable standards that prevents multiple incompatible national frameworks from developing in parallel against a threat that does not respect borders.

Underlying all five reforms is a structural issue: access programs are deployment mechanisms, not governance mechanisms. They determine who uses the technology but do not verify what it does or create ongoing accountability for how it is used. Accountable deployment requires institutional accountability, scoped use authorization, mandatory incident reporting, and tiered access matched to verification depth. These design patterns exist in nuclear materials management, DEA Schedule I research authorization, and export control regimes because they create ongoing accountability rather than one-time verification. Federal procurement requirements should condition access on equivalent frameworks.

5. Confirm CISA Leadership

Finally, CISA (the Cybersecurity and Infrastructure Security Agency), the lead federal agency responsible for coordinating vulnerability disclosure across the civilian government, maintaining the Known Exploited Vulnerabilities catalog, issuing binding operational directives to federal agencies, and convening the cross-sector governance

working groups that this moment requires needs permanent leadership. It is the single point of institutional authority on which every other recommendation in this paper depends. These large structural changes need an engaged leader to invest in such capacity to update disclosure models, stand up new working groups, and issue the revised directives this paper recommends. Without it, every other reform in this section is unlikely to be implemented.

V. ACCESS AND VERIFICATION STANDARDS

The policy reforms above address what governments must do which is create the floor that no individual organization can supply and only governments can impose. But a floor is not a complete governance architecture. The access and verification standards recommendations in this section address what sits above that floor and those are the practices the security community and the laboratories themselves should adopt to ensure that AI vulnerability discovery capabilities are distributed responsibly, validated rigorously, and disclosed in ways that the broader community can act on.

The frontier model providers building these capabilities are acting responsibly and inventing policy in good faith. Anthropic was right to announce publicly that a threshold had been crossed. OpenAI has built a structured access model that others should study. Google is demonstrating that AI can contribute directly to remediation. The argument here is not against any individual program. It is for converging on shared standards that make these programs interoperable and accountable to the broader community.

We propose three standards in their logical order: verification before access, access governed by transparent process, and published data held to rigorous benchmarks.

Standard 1: Independent Verification Before Access Expansion

The security community cannot calibrate its response based solely on vendor-published benchmarks. This is not a criticism of any laboratory's integrity. It is how the community has always operated. When a vendor claims a new detection capability, we test it. When a researcher claims a new exploit technique, we reproduce it. AI-discovered vulnerabilities should be no different.

Independent verification requires structured access for qualified researchers to evaluate these tools against real codebases in realistic environments, with findings published. The questions the community needs answered are specific: How reliable is AI vulnerability chaining across different software environments? What is the actual false positive rate on representative targets? How do these capabilities perform against code outside the training distribution? How quickly is the model's ability to generate working exploits closing on its ability to find vulnerabilities? That last question goes directly to the defender-attacker asymmetry that determines the appropriate pace of response.

Where broad access for testing is not possible, evidence must be shared with sufficient rigor that third parties can assess it: transparent methodology, statistical analysis beyond selected success cases, documented failure modes, cost-per-run economics, and performance variation across target types. These are the standards the community applies to threat intelligence. AI capability announcements justify the same.

Independent verification is not a gate that defenders must wait behind but it is a parallel track that runs alongside structured access, not before it. Verified defenders operating under the access standards described in this section should be able to deploy these capabilities immediately, under scoped authorization and mandatory incident reporting requirements.

Standard 2: Broad, Curated Access Through a Transparent Process

Once a verification framework is in place, defensive access programs should be designed to reach a large set of verified defenders through published criteria and scalable processes. The goal is to maximize positive impact on the overall security environment, which means making these capabilities available not just to the largest technology companies but to the security organizations, managed service providers, and research institutions that serve the hundreds of thousands of organizations that will never be in a launch-day coalition.

Access should be curated, not unlimited. But curation criteria should be published, the application process documented, and decisions made on timelines that reflect the urgency of the problem. The responsible approach when stewarding a capability that addresses a threat affecting everyone is structured access at scale with transparent criteria.

We disclose that Rapid7 was accepted into OpenAI's Trusted Access for Cyber program and remains under consideration at Anthropic. Our view, that transparent and scalable access processes are preferable to opaque and curated ones, stands independent of our access status.

Standard 3: Rigorous Data Standards for Published Capability Claims

Capability announcements of this magnitude should be accompanied by data the community can actually use: transparent methodology, statistical analysis that goes beyond curated successes to report base rates, documented failure modes, cost-per-run economics, false positive rates, and performance variation across target types outside the training distribution.

One data point missing from every laboratory's published materials is whether the model's ability to generate working exploits is advancing at the same rate as its ability to discover vulnerabilities. That question, more than any other, determines whether defenders are structurally ahead, temporarily ahead, or already behind. The laboratories hold the data. They have not published i

VI. THE PATH FORWARD

AI-driven vulnerability discovery is not a future problem. It is a current capability, demonstrated by multiple independent organizations, on an improving trajectory that will not slow down.

The security community cannot wait for perfect standards before acting. But it can act in the right sequence: independent verification before access expansion. Policy reforms (VEP updates, CVE/NVD redesign, binding international standards) must proceed in parallel with industry action, because no individual vendor program can substitute for the governance infrastructure that only governments can build and scale.

The organizations building these capabilities are acting in good faith, and their early programs reflect that. But good intentions and individual programs are not a substitute for common standards backed by accountability structures, independently verified performance data, and binding international coordination. The security community built coordinated disclosure frameworks, vulnerability scoring systems, and information-sharing protocols over 25 years because individual approaches do not scale. AI-driven vulnerability discovery requires the same institutional investment.

The window to get this right is short. The capability is here. We need the programs and standards to match it.

ABOUT RAPID7

Rapid7, Inc. (NASDAQ: RPD) is a global leader in AI-powered managed cybersecurity operations, trusted to advance organizations' cyber resilience. Open and extensible, the Rapid7 Command Platform integrates security data, enriching it with AI, threat intelligence, and 25 years of expertise and innovation to reduce risk and disrupt attackers. As a recognized leader in preemptive managed detection and response (MDR), Rapid7 unifies exposure and detection to transform the cybersecurity operations of more than 11,500 customers worldwide. For more information, visit our website, check out our blog, or follow us on LinkedIn or X.