

Responsible Disclosure in the Age of AI: A Call for Urgent Action

Hon. Melissa Hathaway

Hathaway Global Strategies LLC, Vienna, Virginia, USA

Artificial intelligence is fundamentally reshaping the balance between vulnerability discovery and remediation. Frontier AI models are now capable of autonomously identifying exploitable software vulnerabilities at unprecedented speed and scale. This development exposes decades of accumulated technical debt created by a software industry that prioritized rapid deployment over secure-by-design engineering practices. Drawing on the evolution of software assurance, vulnerability disclosure frameworks, and U.S. cyber policy, this perspective argues that the current moment represents a strategic inflection point for governments, industry, and critical infrastructure operators. The author examines the growing tension between offensive and defensive equities in cyberspace, the emergence of AI-enabled vulnerability discovery capabilities in both the U.S. and China, and the increasing risks posed by unsupported legacy systems and AI-assisted code generation practices. Responsible disclosure can no longer remain a reactive or fragmented process, but must become a coordinated national and international resilience effort involving governments, software vendors, infrastructure operators, and emergency response organizations. The article concludes with an urgent call for accelerated remediation, large-scale patch management coordination, and sustained investment in automated vulnerability repair capabilities before adversaries exploit this rapidly narrowing window of opportunity.

Keywords: responsible disclosure; artificial intelligence; software vulnerabilities; cybersecurity policy; vulnerability disclosure; software assurance; critical infrastructure security; technical debt

Disclaimer: The views expressed in this work are those of the author(s) and do not reflect the official policy or position of their employer(s), the U.S. Military Academy, the Department of War, the U.S. Government, or any subdivisions thereof. © 2026 The Author(s) unless otherwise stated. As an open access journal, The Cyber Defense Review publishes articles under Creative Commons licenses, and authors retain copyright where applicable.



Hon. Melissa Hathaway is a globally recognized thought leader with 35 years of national security experience. She spearheaded the Cyberspace Policy Review for President Obama and led the Comprehensive National Cybersecurity Initiative for President George W. Bush, earning the National Intelligence Reform Medal and the National Intelligence Meritorious Unit Citation. She advises governments, global organizations, and Fortune 500 companies on cybersecurity, enterprise risk management, and emerging technologies. As a strategic advisor and board member, she combines technical and policy expertise to help clients navigate government policy, industry trends, and economic drivers impacting their strategies. She holds a B.A. from The American University, completed graduate studies in international economics and technology transfer policy, and earned a certificate in Information Operations from the U.S. Armed Forces Staff College. She publishes regularly at <https://www.belfercenter.org/person/melissa-hathaway>

For the last four decades, we have allowed the information and communications technology (ICT)—software and hardware industry—to deliver flawed products under the principle: “field it fast and fix it later” (Hathaway 2019). That principle changed in April 2026, when Anthropic and OpenAI released frontier artificial intelligence (AI) models aimed at improving the security posture of every enterprise, critical infrastructure, and government system. These models can autonomously find and exploit vulnerabilities in production software at a depth and speed that previously required experienced human researchers. In fact, they collapsed the time window from sixty days to about four hours. Now a race against the clock begins because finding a vulnerability and fixing it are two entirely different workflows, and the gap between them is where companies find themselves most vulnerable. Now it is time to reconcile the technology debt that we have incurred from these companies that prioritized profit and speed to market over security, privacy, and safety.

The imperative to build trusted products, especially software, dates back to the early 1990s as part of the United States’ Strategic Defense Initiative (SDI). The Trusted Software Methodology was a joint project between General Electric and ATT with the National Security Agency (NSA) acting as an advisor on information security of the SDI (*COMPASS '93, Eighth Annual Conference on Computer Assurance* 1993). 85% of the methodology was simply good software engineering practices to prevent accidental security flaws. The other part of the methodology introduced rigorous operational and developmental constraints to prevent malicious insiders from intentionally injecting backdoors or malicious code into the software and ultimately the end product (*COMPASS '93, Eighth Annual Conference on Computer Assurance* 1993). It was the first notion of delivered uncompromised technology. In 1995, this methodology was adopted at Carnegie Mellon’s Software Engineering Institute’s (SEI) and expanded its Capability Maturity Model (CMM). This harmonization produced the Trusted CMM (T-CMM), aiming to embed security assurances directly into standard software process improvement frameworks (Davis 2013).

NSA defined this as software assurance. “The level of confidence that software is free from vulnerabilities, either intentionally designed into the software or accidentally inserted at any time during its lifecycle, and that the software functions in the intended manner” (Committee on National Security Systems 2015). Despite the methods and best practices being published, industry was not incentivized



Figure 1. Responsible disclosure lifecycle from flaw introduction to remediation and follow-up

to create a good product, nor was it penalized for delivering a bad one. Cybersecurity professionals were trying to defend against a steady increase in adversarial and criminal exploitation of systems, and there was no patching cadence or prioritization that could be used. This inspired David E. Mann and Steven M. Christey to write and present a workshop paper entitled *Towards a Common Enumeration of Vulnerabilities* (Mann and Christey 1999). Of course, this is now the accepted global standard—the Common Vulnerabilities and Exposures (CVE™) identifier and subsequent scoring system.

There are four severity levels that correspond with an expectation for industry to act—meaning patch their product. For example, a critical vulnerability that is being actively exploited with a score between 9 and 10 is expected to be remediated within seven days. A high vulnerability with a score between 7 and 10 is expected to be remediated within 30 days. Of course, companies may request a short extension (e.g., 14 days) if the vulnerability is highly complex or if the fix requires significant architectural changes.

This process became known as *responsible disclosure*—a form of cooperation in which vulnerabilities are reported to the owner of the product/service, allowing the organization the opportunity to diagnose and remedy the vulnerability before detailed vulnerability information is disclosed to third parties or the public. Without a deadline to fix the product and a subsequent deadline to disclose said flaw to the public, some companies would take years to issue patches or ignore vulnerabilities entirely (Epstein 2012).¹ In 2025, it is estimated that over 90% of successful intrusions were the result of software vulnerabilities (Geller 2025). Unfortunately, the tools needed to exploit known vulnerabilities are now just an AI prompt away for most anyone.

This conversation is not new, but the potential consequences are significantly more grave. In 2008, President George W. Bush’s Comprehensive National Cybersecurity Initiative (CNCI) compelled the review of responsible disclosure. At that time, the government was not always eager to have vulnerabilities patched because the vulnerability could be exploited to gather intelligence, or used for military purposes. The benefits or importance of vulnerability disclosure were rarely analyzed,

1. https://cheatsheetseries.owasp.org/cheatsheets/Vulnerability_Disclosure_Cheat_Sheet.html



Figure 2. Equities discussion framework linking intelligence, operational, defensive, and economic security considerations. ©2026 Hathaway Global Strategies, LLC. All rights reserved.

discussed, or prioritized to protect critical infrastructures or our economic interests. This offense vs. defense equities discussion would almost certainly sway toward the intelligence or military interests. During CNCI this conversation began to change (Hathaway 2024). The need to change the policy and practice of disclosure was discussed further and refined in the administration of President Obama.

In April 2014, the U.S. government discussed the process by which it considers disclosing vulnerabilities (Daniel 2014). Key questions included:

- How much is the vulnerable system used in the core internet infrastructure, in other critical infrastructure systems, in the U.S. economy, and/or in national security systems?
- Does the vulnerability, if left un-patched, impose significant risk?
- How much harm could an adversary nation or criminal group do with knowledge of this vulnerability?
- How likely is it that we would know if someone else was exploiting it?
- How badly do we need the intelligence we think we can get from exploiting the vulnerability?
- Could we utilize the vulnerability for a short period of time before we disclose it?
- How likely is it that someone else will discover the vulnerability?
- Can the vulnerability be patched or otherwise mitigated?

In March 2015, the Department of Commerce Internet Policy Task Force (IPTF) requested industry comment to identify substantive cybersecurity issues affecting the digital ecosystem, including responsible disclosure (Department of Commerce, National Telecommunications and Information Administration 2015-04-16). In May 2021, President Biden issued Executive Order 14028, *Improving the Nation's Cybersecurity*, which mandated that federal contractors maintain "responsible disclosure" practices. This included participating in vulnerability disclosure programs (VDPs) and providing a Software Bills of Materials (SBOMs) for products, allowing for better identification of vulnerabilities.

Despite all of these initiatives, companies and governments failed to warn us of the imminent dangers posed by pre-packaged vulnerabilities and exploitable weaknesses. The high-profile cybersecurity

incidents of recent years are symptomatic of the attitude that continues to dominate the development and commercialization of digital technology, in which companies strive to release products as quickly as possible and only worry about security flaws after they have already been deployed. Ultimately, the paradigm of “field it fast, fix it later,” which continues to hold sway in the technology industry, must be overcome (Hathaway 2019). This is even more true now that software developers are using AI to write code—creating applications by instructing AI agents in natural language, rather than writing code manually. Unfortunately, this shortcut is creating an entirely new class of software vulnerabilities.

This is why the bold moves by Anthropic and OpenAI are so important. Now, at least 40 of the largest software and hardware vendors have access to two large models to quickly identify vulnerabilities in their products. They now have the tools to find the flaws in their products and use those same tools to develop the remedy. In fact, Anthropic’s Mythos found critical vulnerabilities in 99% of widely used operating systems and web browsers. In March 2026, Anthropic made clear its commitment to responsible disclosure. (Anthropic 2026). “Anthropic aims to follow the industry standard 90-day disclosure deadline, provide human-reviewed reports with suggested fixes where we can, and pace our submissions to what maintainers can actually absorb.” “We may deviate from this default timeline for various reasons including, for actively exploited critical vulnerabilities and when a finding reflects an ecosystem-wide pattern affecting many projects at once,” among other things.

The U.S. is not alone in these advancements. China’s DeepSeek AI frontier model has achieved similar results (Benincasa 2026). In April 2026, Chinese cyber company 360 Digital Security Group launched an AI-powered “Vulnerability Discovery Agent” that has already uncovered close to 1,000 previously unknown vulnerabilities including in Microsoft’s Office and OpenClaw. This paired with DeepSeek V4 release optimized with Huawei’s Ascend chips makes China both an ally for cleaning up the ICT eco-system and a possible threat to further exploit U.S. vulnerabilities to meet its national interests. It is important to note that in September 2021, China passed a regulation on the “Management of Network Product Security Vulnerabilities” that mandates that businesses operating in China report coding flaws to the government before patching or disclosing them publicly (Cary and Del Rosso 2023). This is not voluntary; it is mandatory.

The strategic inflection point for AI and cybersecurity is now. Over the next 12-24 months, there is an opportunity to buy down the technical debt that we have allowed to accrue for 40 years. We can make the digital ecosystem more secure and safe with software better hardened. This requires the companies that have received the models to take action and be held accountable for remediating their product flaws. These companies have a fiduciary responsibility to fix their product, and if they choose to delay or ignore the evidence, they will likely be liable for subsequent exploitation of their products or services.

As organizations, we need to organize and prepare for a tidal wave of patches to vulnerable products that need to be remediated. Governments need to organize too. They need to work with industry and obtain an understanding of each hardware and software provider’s approach and estimated volume of patches against a timeline in order to pace workflows and surge capacity. These must be mapped across strategic assets, companies, services, etc., to determine: percentage of GDP at risk, which defines our economic stability; the percentage of critical citizen services at risk, which defines our social stability; and the degree of national security risk, which defines our sovereign stability. Leaders must

distinguish between critical, time-sensitive needs and those that are not. This approach needs to be organized and prioritized in the CVE databases with advisories published to communicate the urgent patches and criticality of action. If the CVE program cannot scale to the volume and velocity of patches, a different form of emergency communications should be developed.

Of course, there will be some hardware and software products that have reached end-of-life and are no longer supported, and risk cannot be mitigated. Countless organizations—especially in manufacturing and healthcare—are running legacy or unsupported products and services in their environments and are, therefore, vulnerable to exploitation. These products will have to be replaced as soon as possible. Those costs need to be budgeted for or underwritten through some incentive program, similar to what was done during Y2K. If a methodical approach is taken and a hard deadline is set, it could lead to positive outcomes. In the lead up to the year 2000, businesses and government organizations created special technology teams to ensure that all hardware and software were Y2K-compliant. In some cases, the fix was to replace outdated hardware and/or software, whereas other cases required engineers to replace or rewrite code (National Museum of American History, n.d.). The risks outweighed the costs. We are in the same situation today.

There is a need to accelerate innovations and research and development activities, similar to the Defense Advanced Research Projects Agency (DARPA) Cyber Grand Challenge. DARPA's two-year competition focused on developing AI-enabled software that automatically identified and patched vulnerabilities in the source code that underpins critical infrastructure. This AI Cyber Challenge (AIxCC) competition ended in August 2025 and showed that while still in the early stages, having automated vulnerability repair (AVR) is needed soonest (DARPA 2025-08-08). There are a number of programs emerging, but the tools are not readily available or commercialized, presenting another industrial partnership opportunity to operationalize at speed, and deploy solutions to improve our collective resilience.

We are in a window of extreme vulnerability and we should expect malicious actors, who are now also aware of those newly disclosed vulnerabilities, to exploit unpatched systems and steal sensitive data, hold business hostage through ransomware, knock businesses offline and, in some cases, destroy the IT systems that power hospitals, schools, businesses, and essential services. AI has made exploitation of the vulnerabilities an easy task. Within hours of a discovered vulnerability, the same model can develop the attack path to successfully exploit it. Companies, state and local officials, and emergency responders need to prepare for and update incident response/disaster recovery plans. The Governors of each state should engage the National Guard units who specialize in cybersecurity to determine how quickly they could be activated and deployed to support a state-wide emergency. Collectively, they should conduct tabletop exercises for multiple simultaneous high-severity incidents occurring within the same week, and have playbooks in place for critical incidents (Gadi Evron and Mogull 2026). They should also hold town-hall sessions to raise awareness and strengthen coordination and cooperation of all personnel and citizens. The risk is real and could have material and national consequences.

Bravo to the frontier models who brought responsible disclosure into the limelight, sparking national and international conversations. We need to act fast because the window to prepare for our current and future digital risks is shrinking at a rate that does not favor those who choose to delay.

REFERENCES

- Anthropic. 2026. *Coordinated Vulnerability Disclosure for Claude-Discovered Vulnerabilities*, March 6, 2026. <https://www.anthropic.com/coordinated-vulnerability-disclosure>.
- Benincasa, Eugenio. 2026. “Chinese Firm Claims AI-Driven Bug Discovery Near Claude Mythos Scale.” *Natto Thoughts* (April 22, 2026). <https://www.nattothoughts.com/p/where-is-china-in-ai-driven-vulnerability>.
- Cary, Dakota, and Kristin Del Rosso. 2023. *Sleight of hand: How China Weaponizes Software Vulnerabilities*. Technical report. Atlantic Council, September 6, 2023. <https://www.atlanticcouncil.org/in-depth-research-reports/report/sleight-of-hand-how-china-weaponizes-software-vulnerability/>.
- Committee on National Security Systems. 2015. *CNSSI No. 4009, Committee on National Security Systems (CNSS) Glossary*. Unclassified, April 6, 2015. <https://nsarchive.gwu.edu/document/22385-document-08-committee-national-security>.
- Daniel, Michael. 2014. *Heartbleed: Understanding When We Disclose Cyber Vulnerabilities*. White House Blog, April 28, 2014. <https://obamawhitehouse.archives.gov/blog/2014/04/28/heartbleed-understanding-when-we-disclose-cyber-vulnerabilities>.
- DARPA (Defense Advanced Research Projects Agency). 2025-08-08. *AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense*. <https://www.darpa.mil/news/2025/aixcc-results>.
- Davis, Nooper. 2013. *Secure Software Development Life Cycle Processes*. Technical report. Carnegie Mellon University, Software Engineering Institute. https://www.sei.cmu.edu/documents/430/2013_019_001_297287.pdf.
- Department of Commerce, National Telecommunications and Information Administration. 2015-04-16. *Stakeholder Engagement on Cybersecurity in the Digital Ecosystem*. National Telecommunications and Information Administration. http://www.ntia.doc.gov/files/ntia/publications/cybersecurity_rfc_03132015.pdf.
- Epstein, Jeremy. 2012. *What happens when responsible disclosure fails?* CITP Blog, Princeton University, December 5, 2012. <https://blog.citp.princeton.edu/2012/12/05/what-happens-when-responsible-disclosure-fails/>.
- Gadi Evron, Robert T Lee, and Rich Mogull. 2026. *The “AI Vulnerability Storm”: Building a “Mythos-Ready” Security Program (Version 0.95)*. Cloud Security Alliance, SANS Institute, and OWASP, April 18, 2026. <https://labs.cloudsecurityalliance.org/wp-content/uploads/2026/04/mythosreadyv95.pdf>.
- Geller, Eric. 2025. “Developers Knowingly Push Vulnerable Code, Despite Growing Breach Risk.” *Cybersecurity Dive* (August 15, 2025). <https://www.cybersecuritydive.com/news/software-vulnerabilities-breaches-checkmarx-report/757793/>.
- Hathaway, Melissa. 2019. “Patching our Digital Future is Unsustainable and Dangerous.” In *Governing Cyberspace during a Crisis in Trust*, 80–90. Centre for International Governance Innovation, January 1, 2019. <https://www.cigionline.org/articles/patching-our-digital-future-unsustainable-and-dangerous>.
- Hathaway, Melissa. 2024. “Integrating the IC’s Cyber Security Mission.” *Studies in Intelligence (Special Edition, IRTPA 20 Years On)* 68, no. 5 (December). <https://www.cia.gov/resources/csi/static/9-IRTPA-ODNI-and-cyber-security.pdf>.
- Mann, David E., and Steven M. Christey. 1999. “Towards a Common Enumeration of Vulnerabilities.” In *Proceedings of the 2nd Workshop on Research with Security Vulnerability Databases*. West Lafayette, Indiana, USA, January 21–22, 1999. <https://www.cve.org/Resources/General/Towards-a-Common-Enumeration-of-Vulnerabilities.pdf>.
- COMPASS '93, *Eighth Annual Conference on Computer Assurance*. 1993. Gaithersburg, MD: National Institute of Standards and Technology (NIST), June. <https://doi.org/https://doi.org/10.6028/jres.098.035>.
- National Museum of American History. n.d. *Object Group, Y2K*. <https://americanhistory.si.edu/collections/object-groups/y2k>.