Operation Patchlight

How to leverage advanced AI to give defenders an asymmetric advantage in cybersecurity | **Miles Brundage**

Operation Patchlight

How to leverage advanced AI to give defenders an asymmetric advantage in cybersecurity | Miles Brundage

---

This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

Summary

Open-source code is foundational to America's critical infrastructure; it's a key part of the software running our hospitals, power grids, banks, and more. That software supply chain is dangerously vulnerable, risking harm to Americans and exposing our nation to coercion in the event of a conflict. And our software is only becoming more vulnerable as AI empowers attackers to operate with greater speed, efficacy, and scale.

While AI offers the potential to revolutionize cybersecurity by proactively finding and fixing vulnerabilities, this positive outcome is not guaranteed. Neither open-source code security nor critical infrastructure security is sufficiently well-incentivized today.

We propose a national moonshot to leverage AI to find and fix software vulnerabilities across our economy and ensure that AI becomes an asset and not a liability to America's cybersecurity. This effort has two pillars, each designed to correct a critical market failure, and to make cyberattacks more expensive for bad actors:

- **Fix:** Use cutting-edge AI to find and fix vulnerabilities in open-source code before bad actors can.
- **Empower:** Fund the development and continuous improvement of AI-powered tools that help critical infrastructure defenders work more effectively.

Both pillars will become more effective as AI becomes more capable, and can be easily scaled to different funding levels.

Motivation

Open-source code is foundational for much of our infrastructure in general, including critical infrastructure. According to one assessment, 70% of commercially used code was derived from open-source libraries. But those who maintain this open-source code, and those who are tasked with updating the code in critical infrastructure when open-source code changes are severely under-resourced. This results in unpatched vulnerabilities that regularly cause financial losses, identity theft, and interruption of critical societal functions. Consider the case of patient care at hospitals:

- 61% of hospitals in the US say that ransomware has affected their clinical care, with 17% saying that ransomware has led to serious patient harm.
- 53% of medical equipment at American hospitals contains critical vulnerabilities.
- On average, it takes 491 days to apply critical security updates for hospital equipment.
- The economic costs of cyber incidents in health care alone in the single year 2023 were over $15 billion.

The United States' vulnerable cyber infrastructure exposes us to coercion in the event of a war, since adversaries have the ability to "flip the off switch" on much of our society. While the software supply chain faces known risks, the situation could soon get much worse without proactive efforts to ensure that AI helps defenders more than it helps attackers. Exploitation of code vulnerabilities is on the rise as a risk factor in cybersecurity breaches (though it is not the only factor), and attackers have already begun leveraging AI to increase the scale and efficacy of their attacks.

AI-based vulnerability discovery and patching has enormous potential to improve security, but it may not be deployed at the necessary scale and speed in order to move the needle. Open-source code security is a classic market failure in that the gains from being an attacker exploiting open-source code are concentrated (e.g.,

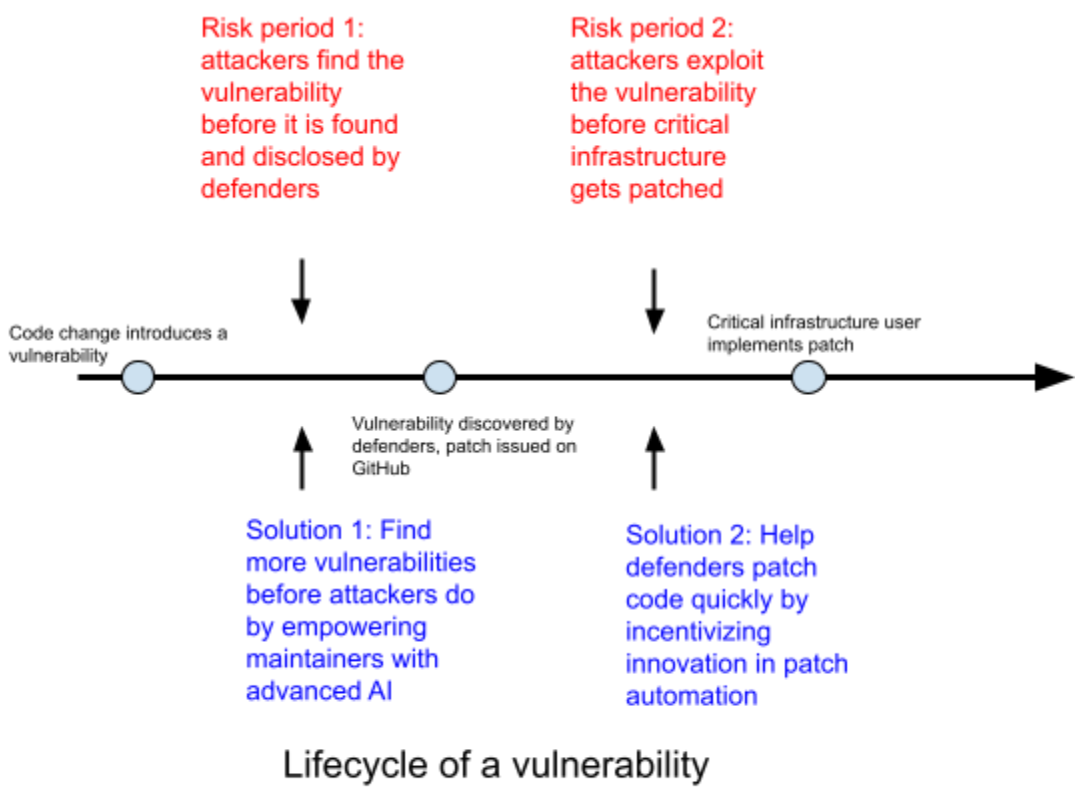
making millions from a single ransomware attack), but the gains from improved cybersecurity are much more diffuse (with many users relying on the same code and counting on others to find the bugs for them). More generally, the under-resourcing of cyberdefense of critical infrastructure means that defenders have limited time to learn about and adopt new technology.

As a growing number of "AI eyes" begin to look for vulnerabilities in open-source code, and exploit the ones they find in our critical infrastructure, we need to turn this common vulnerability into an opportunity. By investing aggressively in the use of AI to fix open-source code faster than bad actors can break it, we can raise the waterline for security across our critical infrastructure. We can go even further by providing under-resourced defenders with the powerful tools they need to easily apply AI-discovered patches and make other security improvements to critical infrastructure code.

Solution

To dramatically improve the security of America's software infrastructure, two steps are needed:

- **Fix:** Use cutting-edge AI to find and fix vulnerabilities in open-source code before bad actors can.
- **Empower:** Fund the development and continuous improvement of AI-powered tools that help critical infrastructure defenders work more effectively.



Lifecycle of a vulnerability

In combination, these would give defenders of critical infrastructure a more reliable software supply chain to draw on and an always-on AI security assistant that works 24/7 and constantly improves over time. Given the scale and cost of the cybersecurity risks we face, the return on investment would be very positive if this made even a small dent in critical infrastructure preparedness.

The US government should therefore launch an effort to coordinate AI labs, cyberdefense actors, industry and philanthropy to secure open-source code and critical infrastructure. This initiative will benefit from the White House's convening power and bring American AI innovation to bear directly on the critical challenges above. The specific version of the proposal described below would involve $2.4 billion in direct funding over 3 years and additional commitments from industry. The funding for this effort could be sourced from a mix of government, philanthropy and industry, and could be scaled up or down while maintaining the same basic structure.

Specifically, the effort would:

- Convene frontier AI labs and provide them incentives to give cyber defenders early access to AI models, and to provide free/low-cost compute/access to cyber defenders securing critical infrastructure.
- Fund initiatives working on open-source code security to enable them to leverage that early access to find and patch vulnerabilities.
- Provide catalytic funding to the development of AI-powered tools for critical infrastructure defenders.

If successful, this effort would create an altered capability and cost environment more favorable to defense efforts, and seed an ecosystem of increasingly powerful defensive tools that could be further developed by industry.

Implementation

"Fix": Use cutting-edge AI to find and fix vulnerabilities in open-source code before bad actors can.

First, the federal government should secure commitments from American AI companies to provide early, free or low-cost access to each generation of frontier AI systems to key open-source code maintainers.

- This should be implemented flexibly; some companies may be better positioned to provide API credits, and others might be better positioned to provide a certain number of free premium accounts on their consumer service (e.g., each open-source project might receive some number of Claude Max, Gemini Advanced, or ChatGPT Pro accounts). Others might want to take recommendations for codebases to focus on using prototype AI systems that have not yet been productized. Each of these would help to put the AI industry's thumbs on the scale in favor of defense.
- In addition to using its convening power, the federal government could tie these actions to "carrots" like expedited datacenter siting, just as it is already doing in order to incentivize stronger cybersecurity at the frontier AI companies themselves as well as the provision of AI for national security purposes.

Low-cost and early access on its own is not enough. Many open-source software libraries are so under-resourced that they would not have the capacity to leverage that access effectively. Thus, we also recommend that the federal government and philanthropic institutions provide funding for non-profit organizations focused on open-source code security:

- Funding would cover the costs associated with securing especially critical "load-bearing" open-source projects. Particular attention should be paid to open-source software that is heavily used in critical infrastructure.
- In-scope activities would include using AI APIs to conduct "high-compute" vulnerability discovery on select open-source codebases, generation of patches through similar means, and payment for the labor costs associated with codebase prioritization, manual vulnerability and patch inspection, and conducting outreach to open-source maintainers to inform them about the program.
- As one example, the Alpha-Omega initiative at the Open Source Security Foundation (OpenSSF) has a track record of improving open-source code security, but it is underfunded relative to the scale of the challenge and opportunity. Historical funding for this initiative is in the low millions, which is tiny compared to the billions in damage that have sometimes been caused by individual vulnerabilities in open-source code.
- A commitment of $100 million for the remainder of 2025, followed by $300 million for 2026 and $1 billion in 2027, would help attract new talent and energy into open-source code security efforts.

"Empower": Fund the development and continuous improvement of AI-powered tools that help critical infrastructure defenders work more effectively.

Finding vulnerabilities before attackers is not enough. They need to be patched before they're exploited, and indeed, many known vulnerabilities are constantly being exploited because defenders of critical infrastructure are overwhelmed and under-resourced (often being just solo administrators or small teams). Despite

being numerous and vital to our nation, these defenders are not the primary customer base for Silicon Valley startups building cybersecurity tooling, precisely because of their under-resourcing and dispersed nature (unlike, e.g., Fortune 500 companies that buy many licenses for enterprise software).

To equip critical infrastructure defenders with the necessary capabilities, we recommend the following:

- The federal government should invest $1 billion in rapidly issued grants to companies or non-profits building AI-powered tooling for critical infrastructure defenders. The aim is to provide an upfront commitment of capital that increases attention to and innovation in this market, making it more likely that private capital will subsequently flow to and help scale up this ecosystem. This dollar amount could be scaled up or down but is intended to allow for several bets to be placed at a scale that would attract attention in a very active and well-funded AI innovation landscape.
- The federal government should also secure commitments from frontier AI companies to provide at least $1 billion per year for five years in free or reduced-cost AI capabilities to augment software that is in active use by critical infrastructure defenders. Critically, funding recipients should be eligible regardless of whether they were recipients of the initial funding described above (to avoid premature market consolidation), and recipients should be accredited via a lightweight process administered directly or indirectly by the Cybersecurity and Infrastructure Security Agency (CISA). $1 billion is intended to approximate significant volumes of AI usage, at current prices, though the actual costs to the companies would be much less than $1 billion, since they sell AI “tokens” at a profit.

Together, these recommendations would make it possible for every system administrator at, for example, an American hospital or power plant to have an always-on, eager AI assistant that helps them incorporate new software patches to their network. This software would have the following properties:

- Uses the latest AI models via an API or open-source models on-site.
- Functions like a virtual colleague that works smoothly on email, Slack, GitHub, and other platforms, rather than requiring a new learning curve.

- Proactively suggests edits to software in order to accommodate any clashes created by new patches, and produces appropriate documentation to make patching as easy as possible.

Such AI tools could help address the current backlog of known vulnerabilities and make sure that the coming wave of AI-enabled vulnerability discovery is helpful for defenders, rather than just growing the backlog.

While our primary focus here is on fixing known vulnerabilities, there are other ways in which an always-on AI assistant could augment the capacity of critical infrastructure defenders, and some portion of this investment should also go towards wider exploration of these possibilities. Many breaches result from lack of adoption of key security interventions such as multi-factor authentication. AI tooling could also be used to help scale communications with staff about the importance of such safeguards in a way that is more tailored, effective, and up-to-date than yearly training videos or boilerplate emails. An always-on AI security assistant could also potentially help in areas such as intrusion detection and incident response.

The goal: Cyberdefense dominance

Both of these moonshot pillars will naturally become more effective as AI capabilities mature, and will infuse more AI into our critical infrastructure with each cost reduction in AI APIs. Additionally, each can be funded in a distributed fashion, and can scale smoothly — further increasing the cost of a successful attack. With more investment, we will asymptotically approach cyberdefense dominance, locking low-end attackers out of the market. Eventually, through a combination of this proposal and others such as The Great Refactor, we can even lock high-end attackers out of our critical infrastructure.

The greatest risk to this moonshot is not technical failure, but a failure of ambition. Half-measures will be lost in the noise of a rapidly changing software ecosystem and a noisy AI startup landscape. The maintainers and defenders on the front lines have many competing priorities; only a vigorous, well-funded national effort can provide the momentum needed to truly shift the cybersecurity landscape. By pursuing this moonshot with conviction and drawing on our best humans and best AIs, we can safeguard America's critical infrastructure for the years ahead.

Further resources

- OpenSSF, Alpha-Omega Project, 2024.

  We recommend doubling down on this program, which is well-aligned with the goals of the moonshot described here. Their website includes information on grants they have given out and their approach to open-source project prioritization.

- Sean Heelan, How I Used o3 to Find CVE-2025-37899, a Remote Zero-day Vulnerability in the Linux Kernel's SMB Implementation, 2025.

  A recent blog post describing an example of applying OpenAI's o3 model to vulnerability discovery. The post captures the current moment in cybersecurity, during which AI is rapidly moving from a curiosity to an essential ingredient in attack and defense.

- XBOW, How XBOW Did It: AI-driven Penetration Testing and the Discovery of Open Source Vulnerabilities, 2025.

  A recent blog post describing the company XBOW's approach to AI-driven penetration testing, which has surfaced many vulnerabilities in open-source code. The post, which came out just weeks after the one above, again shows the fast pace of AI's transformation of cybersecurity. It also illustrates the fact that many vulnerabilities in open-source code likely remain to be discovered, and that if urgent action isn't taken to further invest in defensive efforts like this, malicious actors will deploy very similar technology soon and exploit the vulnerabilities they find.

- ARPA-H, UPGRADE Proposers Day Presentation, 2024.

  An example of current government interest in automating software patching, as well as illuminating statistics on the severity of cybersecurity challenges in the healthcare sector.

- Cybersecurity and Infrastructure Security Agency, CISA Performance Goals Adoption Report, 2025.

  An authoritative recent overview of the current state of American cybersecurity, by the Cybersecurity and Infrastructure Security Agency.