THE DEATH OF RESPONSIBLE DISCLOSURE

How AI Made Vulnerability Discovery Free — And Left Defenders Alone With The Bill

A Crucible Intelligence Report
April 2026
thecrucible.systems

Executive Summary

Responsible disclosure — the institutional framework governing how security vulnerabilities are found, verified, reported and patched — is structurally dead. Not dying. Not under pressure. Dead.

The evidence is already in the numbers: 32% of vulnerabilities are exploited on or before the day the CVE is issued. Median time from disclosure to exploitation has collapsed from 8.5 days to 5.0 days. These are not predictions about what AI will do. They are proof that the framework failed before AI acceleration even arrived.

What AI has done is make the failure visible — and irreversible. Nicholas Carlini, a research scientist at Anthropic, used Claude Code to find hundreds of remotely exploitable Linux kernel vulnerabilities including one hidden for 23 years. He cannot report most of them. Not because the bugs are not real. Because human validation is the bottleneck and he will not send maintainers unverified findings.

Attackers running the identical script have no such constraint.

> **The core asymmetry:** The validation bottleneck only binds defenders. Responsible disclosure was built for human-speed research. AI has made discovery essentially free. The framework was designed for a world that no longer exists.

This paper presents five findings derived from structured epistemic analysis using the Crucible decision intelligence platform. It concludes with a named replacement framework: Post-Exploitation Response Coordination — and what it will take to build it.

Section 1: The Discovery Inversion

For decades, finding security vulnerabilities in complex codebases was the hard part. It required deep domain expertise, manual code review, and significant time investment. The bottleneck was discovery.

That constraint no longer exists.

Nicholas Carlini used a 12-line bash script to point Claude Code at the Linux kernel source and ask it to find vulnerabilities. The script looped over every file in the kernel, asking Claude to focus on each one in turn. The result: hundreds of potential crashes and remotely exploitable heap buffer overflows — including CVE-2026-5133b61, a bug in the NFS driver introduced in March 2003 and undiscovered for 23 years.

> *"I have never found one of these in my life before. This is very, very, very hard to do. With these language models, I have a bunch."*

— Nicholas Carlini, [un]prompted 2026

The bug itself required understanding intricate details of NFS protocol interaction across two cooperating clients. It was not a simple pattern match. Claude Code found it in minutes. It sat undiscovered through thousands of human code reviews for over two decades.

Carlini found five bugs that have been reported and patched. He is holding several hundred more. Not because they are unimportant. Because the validation step — confirming the bug is

real and exploitable before reporting — requires human expertise that cannot be scaled to match AI discovery speed.

The inversion: Discovery used to be the constraint. Validation is now the constraint. The bottleneck has flipped — and it only binds the people trying to fix the problem, not the people trying to exploit it.

Section 2: The Asymmetry That Cannot Be Patched

Responsible disclosure rests on a single assumption: defenders get a window. Find the bug, validate it, report it, patch it — before attackers find and weaponise it. The whole framework is built around that sequence holding.

It does not hold.

The validation constraint is not symmetrical. When Carlini runs Claude Code and finds a potential bug, he must verify it before reporting. He will not send maintainers unverified output. That is the responsible disclosure contract.

An attacker running the identical script has no such requirement. They do not need to confirm the bug is real before attempting exploitation. They probe directly. The validation step simply does not exist in the attack chain.

This asymmetry is not a gap that better tooling will close. It is structural. It is baked into the definition of responsible disclosure itself. The framework requires defenders to verify. It places no equivalent burden on attackers.

Industry leaders at RSAC 2026 named this directly:

> *"Because of the asymmetry in the cyber domain, where one person on offense can create work for millions of defenders, speed leverages that asymmetry. In the near term, there's an advantage to the attackers as they start to use models and agents to do a lot of the offense."*

— Kevin Mandia, Founder, Armadin (RSAC 2026)

The structural problem: Responsible disclosure requires defenders to validate. AI makes discovery free for everyone. The validation burden only falls on the side that is

trying to help. This is not a fixable edge case — it is the framework working as designed, in a world it was not designed for.

Section 3: The Infrastructure Was Already Failing

Before accounting for AI acceleration, responsible disclosure was already under pressure. The numbers from 2025 are unambiguous.

The Timeline Has Collapsed

- 32% of vulnerabilities are exploited on or before the day the CVE is issued
- Median exploitation time has dropped from 8.5 days to 5.0 days
- Over 40,000 CVEs were reported in 2024 alone — a 38% increase from 2023
- AI systems can generate working CVE exploits in 10-15 minutes at approximately $1.00 per exploit
- Attackers can operationalise more than 130 new CVEs daily at scale

These figures mean the validation window — the time between responsible disclosure and exploitation — does not exist in a meaningful percentage of cases. Defenders are not losing the race. They are running in a race that is already over.

The Maintainer Bottleneck

Even when bugs are found and validated, the patch pipeline strains under volume. The Linux kernel has over 75% of files with multiple maintainers — but expertise is already distributed to capacity limits. Subsystem maintainers cannot afford to wait for reviewers, with response expectations ranging from two days to several weeks depending on subsystem.

When AI produces hundreds of validated findings simultaneously, the triage queue becomes structurally unworkable. The social and technical dynamics of kernel patching — developer responsibility, subsystem structure, backport difficulty — were not designed for machine-speed input.

The proof: Responsible disclosure did not fail because of AI. AI made the failure measurable. 32% exploitation on day zero is the system working as designed — and that design is obsolete.

Section 4: The Wave That Is Already Here

Nicholas Carlini's findings used Claude Opus 4.6. He tested older models and found that Opus 4.1 and Sonnet 4.5 could find only a small fraction of what Opus 4.6 discovered. The capability curve is steep and recent.

This is not a future threat. The wave described at RSAC 2026 is present tense.

> *"Foundation model companies are sitting on thousands of bugs discovered through AI-assisted analysis that they lack the capacity to verify or patch. The exploit discovery has gone exponential."*

— Alex Stamos, CSO Corridor (RSAC 2026)

The democratisation problem compounds this. When open-source models reach current capability levels, the barrier to sophisticated vulnerability research disappears entirely. Elite capability becomes accessible to anyone with a laptop and motivation.

This creates a second-order problem beyond individual vulnerabilities: the entire coordination infrastructure of responsible disclosure begins to strain when discovery outpaces human triage capacity. The framework assumes a manageable flow of findings. It was not designed for simultaneous hundreds-of-findings input from a single researcher's overnight run.

The trajectory: AI capability for vulnerability discovery doubles approximately every four months. The validation capacity of human maintainers does not. The gap is not closing — it is widening at an accelerating rate.

Section 5: Post-Exploitation Response Coordination

If responsible disclosure is structurally obsolete, the question is not how to fix it. The question is what replaces it.

The answer emerging from the analysis is a named framework: Post-Exploitation Response Coordination. It accepts the premise that AI-generated vulnerabilities will be exploited before

they can be validated or disclosed — and rebuilds coordination infrastructure around rapid response rather than prevention.

The objective shifts. Not perfect prevention. Not closing every CVE before exploitation. Instead: achieving detection, blocking, and recovery speed that matches the weaponisation pace of attackers.

What the Framework Requires

Real-Time Attack Detection Networks:

- Automated alert triage operating at machine speed
- Autonomous incident response that does not wait for human review
- Continuous security operations that treat exploitation as expected, not exceptional

Machine-Speed Patch Generation:

- AI systems that reason about code weaknesses and generate reliable patches
- Automated patch verification and regression testing
- Deployment infrastructure that can push fixes faster than exploitation can propagate

Coordinated Response Infrastructure:

- Continuous asset identification and attack surface mapping
- Emergency patch SLAs measured in hours, not days or weeks
- Focused threat hunting that assumes breach rather than prevention
- Rapid recovery through established backup and incident response playbooks

What Would Make This Wrong

Post-Exploitation Response Coordination rests on the assumption that the validation window is gone. It would be wrong if:

- AI validation achieves sub-5% false positive rates — making preventive disclosure viable again
- Attack-defence parity emerges through equivalent AI capability on the defensive side
- Regulatory frameworks mandate validation-dependent disclosure regardless of economic feasibility
- AI cascading failures across critical infrastructure prove impossible to contain through rapid response

The honest assessment: The framework is not a solution. It is an adaptation to a world where the solution — closing vulnerabilities before exploitation — is no longer achievable at the pace required. The goal is resilience, not prevention.

What This Means for Future Work

The shift from discovery-limited to validation-limited security has implications that extend well beyond the Linux kernel or responsible disclosure specifically.

For Security Researchers

The individual researcher's advantage has inverted. The skill that mattered — finding bugs — is now essentially free. The skill that matters now is triage judgment: distinguishing which of hundreds of AI-generated findings are real, exploitable, and worth the coordination cost of reporting. That judgment cannot be automated at sufficient accuracy yet. It is the new scarce resource.

For Software Maintainers

The maintainer pipeline was designed for human-speed input. It will need structural redesign to accept machine-speed findings. This means automated triage layers, AI-assisted validation tools, and prioritisation frameworks that can handle volume without burning out the human experts at the top of the chain. The social dynamics of open-source maintenance — trust, reputation, relationship — do not disappear. They become more important as the signal-to-noise ratio in incoming reports deteriorates.

For Policy and Regulatory Bodies

Current disclosure frameworks are written for human-speed research. Regulation that mandates validation timelines, disclosure windows, or coordination requirements will need to be revisited against the reality that those timelines no longer reflect how discovery actually works. The 90-day disclosure window made sense when finding a bug took weeks. It is incoherent when finding hundreds of bugs takes an overnight run.

For AI Developers

The tools that enable this — Claude Code and its equivalents — are not neutral. The same capability that lets Carlini find bugs for responsible disclosure lets attackers find bugs for

exploitation. The asymmetry is a product design problem as much as an institutional one. How AI labs choose to deploy, restrict, and monitor use of vulnerability-discovery capabilities will shape how the transition to Post-Exploitation Response Coordination unfolds.

How Crucible Can Help

This white paper was produced using Crucible — a decision intelligence platform. Every finding in this report emerged from structured epistemic pressure applied to primary source material, not from opinion or secondary commentary.

The Crucible methodology is built for exactly the class of problem this paper addresses: decisions and frameworks where the stakes are high, the evidence is ambiguous, and conventional analysis produces generic conclusions.

For Security Teams

Crucible can pressure-test your organisation's current security posture against the Post-Exploitation Response Coordination framework. Not as a product recommendation engine — as an analytical platform that finds where your assumptions break before your attackers do.

- Identify which elements of your current disclosure and response processes are built for human-speed assumptions
- Find the structural gaps in your validation pipeline before AI-assisted attackers find them operationally
- Test your incident response framework against machine-speed exploitation scenarios

For Policy and Research Teams

The responsible disclosure replacement problem is an institutional design challenge. Crucible is well-suited to pressure-testing proposed frameworks, identifying hidden assumptions in policy proposals, and finding what conditions would need to hold for a given approach to work.

- Evaluate proposed disclosure frameworks against the asymmetry constraint
- Identify regulatory approaches that would survive machine-speed discovery
- Find the irreducible human judgment requirements that automation cannot replace

For AI Developers and Labs

The capability-responsibility gap — between what AI can discover and what institutions can handle — is a design problem. Crucible can help think through deployment frameworks, disclosure policies, and coordination mechanisms that are coherent with the actual pace of AI capability growth rather than the pace institutions were built for.

---

thecrucible.systems

Decision intelligence for decisions that matter.

thecrucible.systems